ABSTRACT

Finishing is rate limiting for genome projects, and improvements in the efficiency of complete genome-sequence compilation will require improved protocols for gap closure. Here we report a novel approach for extending shotgun contigs and closing gaps that we termed PCR-assisted contig extension (PACE). PACE depends on the capture of rare mismatched interactions that occur between arbitrary primers and template DNA of unknown sequence, even under highly stringent conditions, by means of elevated PCR-cycle repetition and the use of specific anchoring primers corresponding to adjacent regions of known sequence. The use of two or more cloning systems and libraries with different insert sizes tends to reduce the number of gaps for which no cloned DNA is available. Gaps are most frequently closed by combinatorial PCR with primers derived from contig ends (1). Nevertheless, this is ineffective for larger gaps and often generates nonspecific fragments. Alternative strategies for gap closure include subtractive hybridization (2), direct sequencing of genomic DNA (3,4), and physical mapping strategies based on macro-restriction maps (5). However, all of these are technically complex and time consuming.

In the course of completing the 5-Mb genome sequence of the free-living bacterium Chromobacterium violaceum, we used a novel approach for gap closure that we found simplified and accelerated the closure of the genome. The technique involves the generation of stepwise extensions from the ends of contigs by PCR until the closure of individual gaps is achieved. Several PCR methods have been described for the isolation of DNA segments adjacent to known sequences that could, in principle, have been exploited for such a strategy. The majority of these have been utilized for the isolation of 5' flanking genomic regions of cDNA clones to clone insertion-tagged genes and to isolate the extremities of the inserts of large clones such as those in YACs (6–13). However, none of these methods have been previously evaluated as a strategy for bacterial genome closure. The majority of the relevant PCR methodologies require either processing of the template DNA before amplification or post-amplification purification and cloning steps, thus limiting their applicability within the high-throughput environment of a genome-sequencing project.

However, we selected one method in which these constraints do not apply that had been originally developed for the isolation of the ends of YAC inserts (11). We have termed the stepwise application of this methodology for gap closure in bacterial genome projects PCR-assisted contig extension (PACE).

MATERIALS AND METHODS

Primers

A set of 96 primers was selected from stocks that remained from previous sequencing projects with no reference to their precise sequence, but with a preference for primers with an average of 19 bp and a 60°C annealing temperature. The primers (5 μM) were re-suspended in water and arrayed on a 96-well plate. Pairs of outward-facing specific nested primers were designed approximately 140 bp from contig ends and 40 bp apart from each other.

INTRODUCTION

Bacterial genome projects typically accrue most sequence data through high-throughput shotgun sequencing. The resulting shotgun draft is then converted into a complete representation of the organism’s DNA through a more laborious completion phase. This phase is devoted mainly to the generation of additional sequence data to close gaps caused by statistical cloning fluctuations and the existence of sequences refractory to the cloning system used. The use of two or more cloning systems and libraries with different insert sizes tends to reduce the number of gaps for which no cloned DNA is available. Gaps are most frequently closed by combinatorial PCR with primers derived from contig ends (1). Nevertheless, this is ineffective for larger gaps and often generates nonspecific fragments. Alternative strategies for gap closure include subtractive hybridization (2), direct sequencing of genomic DNA (3,4), and physical mapping strategies based on macro-restriction maps (5). However, all of these are technically complex and time consuming.

In the course of completing the 5-Mb genome sequence of the free-living bacterium Chromobacterium violaceum, we used a novel approach for gap closure that we found simplified and accelerated the closure of the genome. The technique involves the generation of stepwise extensions from the ends of contigs by PCR until the closure of individual gaps is achieved. Several PCR methods have been described for the isolation of DNA segments adjacent to known sequences that could, in principle, have been exploited for such a strategy. The majority of these have been utilized for the isolation of 5' flanking genomic regions of cDNA clones to clone insertion-tagged genes and to isolate the extremities of the inserts of large clones such as those in YACs (6–13). However, none of these methods have been previously evaluated as a strategy for bacterial genome closure. The majority of the relevant PCR methodologies require either processing of the template DNA before amplification or post-amplification purification and cloning steps, thus limiting their applicability within the high-throughput environment of a genome-sequencing project.

However, we selected one method in which these constraints do not apply that had been originally developed for the isolation of the ends of YAC inserts (11). We have termed the stepwise application of this methodology for gap closure in bacterial genome projects PCR-assisted contig extension (PACE).
PCR Amplifications

PACE was developed as a two-step PCR strategy with nested primers derived from contig ends as depicted in Figure 1. Both reactions were performed on 96-well plates in a final volume of 20 μL. Reaction mixes contained 200 μM dNTPs, 0.5 μM each primer, 1.5 mM MgCl2, Taq DNA polymerase (Invitrogen, Carlsbad, CA, USA) and appropriate buffer. In step 1, 80 ng genomic DNA and 1.5 U Taq DNA polymerase were added per reaction. In step 2, a 1:100 dilution of the step 1 reaction products was transferred to a second 96-well plate and used as a template to which was also added the same arbitrary primer used in the step 1 reaction and 1 U Taq DNA polymerase. Step 1 consisted of 60 cycles at 94°C for 10 s, 55°C for 15 s, and 72°C for 2 min, with an initial denaturation at 94°C for 2 min and a final extension at 72°C for 3 min. Step 2 was identical but with an annealing temperature of 60°C and 40 cycles.

PACE Product Analysis

Three microliters of each reaction were loaded onto 1% agarose gels with step 1 and step 2 products in adjacent lanes. Reaction products that consisted of a single detectable band were purified with the QIAquick™ PCR Purification Kit (Qiagen, Valencia, CA, USA) and sequenced directly with the same specific primer used for amplification on an ABI PRISM® 3100 DNA sequencer (Applied Biosystems, Foster City, CA, USA). High-quality sequences with more than 300 bp and Phred quality greater than 20 were analyzed for specificity. The specificity of the PACE products was verified by BLASTN analysis against the available genome assembly. All contig extensions and gap closures were confirmed by specific PCR, followed by the direct sequencing of the PCR products.

RESULTS AND DISCUSSION

The PACE technique depends on rare mismatched primer-template interactions to be captured through elevated PCR-cycle repetition (Figure 1) and was developed in the context of the completion of the C. violaceum genome. Figure 2 shows the products generated from a typical set of PACE reactions that are all derived using the same pair of nested primers. A total of 43 visible products were obtained in step 1 reactions, and 71 products were obtained in step 2 reactions. Single amplified fragments were observed in 16 and 25 of the step 1 and step 2 reactions, respectively. Sequences from these products were compared to the C. violaceum genome and

![Figure 1](image1.png)

Figure 1. A schematic outline of the PACE reaction.

![Figure 2](image2.png)

Figure 2. A 1% ethidium bromide-stained agarose gel containing PACE reaction products derived for a single C. violaceum contig using the same pair of nested primers for each set of step 1 and step 2 reactions. The numbers at the top and the letters (A–H) on the left-hand side represent the position in the microplate. Step 1 and step 2 reactions using the same arbitrary primer have been run in adjacent lanes. The asterisks represent products selected for sequencing. M, 100-bp size marker ladder (MBI Fermentas, St. Leon-Rot, Germany).
classified as authentic contig extensions when they matched with the extremity of the contig from which the pair of specific nested primers was derived. Of the 43 single fragments, 37 were selected for sequencing based on their size and intensity, and 21 generated high-quality sequence (300 bp with Phred > 20). All high-quality sequences were shown to be specific to the expected contig.

PACE was used to extend 22 contig ends from the *C. violaceum* genome. Table 1 summarizes the results. From the 22 sets of PACE reactions, a total of 1888 individual reactions resulted in visible products that ranged in size from 0.4 to 3 kb, with an average size of 1 kb. A total of 688 products were obtained from step 1 reactions, with an average of 31.3 products per set of reactions, and 1200 were obtained from step 2 reactions, with an average of 54.5 products per set of reactions. The step 2 reactions gave a higher percentage of single-band products (35.6%) than the step 1 reactions (24.8%).

A total of 103 single-band step 1 products and 311 step 2 products were selected for sequencing. A higher percentage of high-quality sequences were obtained from step 2 products (41.5%) than step 1 products (20.4%). The percentage of specific product was also much higher for the step 2 products (97.7%) than for the step 1 products (52.4%). The sequencing of single PACE products allowed for the immediate closure of 15 real gaps in the *C. violaceum* genome sequence because of their relatively small size. In the remainder of the cases, additional sets of PACE reactions were undertaken until closure was achieved.

No correlation between amplification efficiency/specificity and primer length or base composition was observed (data not shown). The large number of correct PACE products per reaction suggests that the use of 32 arbitrary primers was sufficient to ensure that at least one specific fragment of 1–3 kb per contig. Additionally, we recommend the sequencing of only step 2

Table 1. Summary of Results Obtained with PACE in Extending *C. violaceum* Shotgun Contigs

<table>
<thead>
<tr>
<th>Contig&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Amplification&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Single Products&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Sequences&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Specific Products&lt;sup&gt;e&lt;/sup&gt;</th>
<th>Largest Single Product&lt;sup&gt;f&lt;/sup&gt; (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>step 1</td>
<td>step 2</td>
<td>step 1</td>
<td>step 2</td>
<td>step 1</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>48</td>
<td>6</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>72</td>
<td>8</td>
<td>29</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>69</td>
<td>6</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>76</td>
<td>12</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>40</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
<td>71</td>
<td>16</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>58</td>
<td>2</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>46</td>
<td>7</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>54</td>
<td>11</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
<td>72</td>
<td>15</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>56</td>
<td>3</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>33</td>
<td>45</td>
<td>10</td>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>55</td>
<td>59</td>
<td>14</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>39</td>
<td>56</td>
<td>10</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>26</td>
<td>44</td>
<td>9</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>46</td>
<td>55</td>
<td>6</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>53</td>
<td>49</td>
<td>11</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>49</td>
<td>8</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>15</td>
<td>56</td>
<td>1</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>36</td>
<td>53</td>
<td>6</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>28</td>
<td>38</td>
<td>3</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>14</td>
<td>34</td>
<td>6</td>
<td>17</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>Contig, number of contig ends.
<sup>b</sup>Amplification, number of reactions with detectable products.
<sup>c</sup>Single product, number of reactions that resulted in a single detectable fragment.
<sup>d</sup>Sequences, number of sequences analyzed.
<sup>e</sup>Specific products, number of specific products judged to be authentic by BLASTN.
<sup>f</sup>Largest single product, the size of the largest single product from each PACE reaction in step 2.
reaction products and suggest the analysis of at least two of these for sequence confirmation.

The PACE methodology proved especially useful for extending the multi-copy ribosomal operons. When we started using PACE to close the *C. violaceum* genome assembly, seven apparently identical copies of rRNA operons composed of one copy of the 5S, 23S, and 16S genes had been identified. Although seven contigs ended with 5S rRNA gene sequence, only three ended with the 16S rRNA gene. We employed PACE to generate sequence flanking for the remainder of the 16S rRNA genes using a single pair of nested primers. We generated a total of 21 step 1 products and 59 step 2 products of which 13 and 32, respectively, consisted of single DNA fragments. From these fragments, 35 high-quality sequences were obtained. Eighty-nine percent (8/9) of step 1 products and 100% (26/26) of step 2 products were found to be specific and constituted five different flanking regions for the 16S rRNA gene and the three already known, resulting in a total of eight rRNA operons. We were thus required to undertake a 5S rRNA gene-based PACE reaction to complete the final operon that had been discovered as a result of the 16S PACE protocol.

To address the optimal timing for the employment of PACE in the course of a genome project, we used the data derived from the *C. violaceum* genome assembly. We calculated the total number of PACE reactions needed to close the genome by dividing the uncovered fraction of the genome into base pairs divided by the expected average size of a PACE extension (1 kb). We estimated that a PACE-mediated contig extension is 24.6 times more expensive than a single shotgun read, although this will vary with primer and sequence cost. This estimate was based on the use of 32 arbitrary amplifications and the sequencing of two PCR products in both directions. While estimating the cost of the PACE reactions, we took into account the synthesis of specific step 1 and step 2 primers, *Taq* DNA polymerase, tips, 96-well plates, purification of PCR products, and sequencing of single products. For the shotgun-sequencing cost estimate, we considered the number of reads used in the progressive assembly of the *C. violaceum* genome and a read cost of US $3.00.

Figure 3 presents the total genome-sequencing costs when applying PACE at different stages of shotgun sequencing. If it is applied at 3-fold genome coverage, then 1920 sets of PACE reactions are required for closure, thus ele-

![Figure 3. Relative overall cost of a complete 5-Mb genome sequence generated using PACE, following different levels of shotgun coverage.](image-url)
vating the total cost. At 6-fold coverage, the number of PACE reactions is drastically reduced (222 PACE reactions). The lowest cost for the project is obtained at approximately 5-fold genome coverage, when 345 PACE reactions are needed to close the genome. However, the continuation of shotgun sequencing until approximately 8-fold coverage does not significantly increase the total project cost and will also considerably improve the overall quality of the sequence (Figure 3). At this point, the number of PACE reactions required is 123 of a 5-Mb genome.

The importance of PACE is that, at a predetermined point in the shotgun sequencing, primers can be generated and contig extensions obtained with no requirement for a one-by-one analysis of the gaps. For each set of reactions, the two longest single-step 2 fragments (Table 1, column 10) can be sequenced, and the procedure can be repeated until the gap is closed. Using successive PACE reactions, it is possible to close gaps in bacterial genomes in a stepwise fashion regardless of their size. Nevertheless, the larger the gap is, the more rounds of PACE that will be required. With the availability of an in-house DNA synthesizer, this procedure could be almost entirely automated and would be expected to result in genome closure in a couple of days. The approach is also applicable in other instances in which flanking sequences are required, such as in identifying the position of insertional mutation elements (10, 12, 13, 15–17). Preliminary results using the human genome DNA as a template and the current PACE format revealed a low percentage of single products (8.3% and 12.5% for step 1 and 2, respectively) in comparison with the data generated for C. violaceum [24.8% and 35.6% for step 1 and 2, respectively (Table 1)]. Of the single fragments selected for sequencing, 16.7% generated high-quality sequence; however, none of them were specific for the respective genome regions. These results suggest that the current PACE format is inappropriate for gap closure in more complex genomes, for which it will require additional modifications. Further experimentation will indicate the upper level of genome size where the current PACE protocol remains effective.

ACKNOWLEDGMENTS

The work described here was undertaken within the context of the Brazilian National Genome Program, a consortium funded in December 2000 by the Ministério de Ciência e Tecnologia (MCT) through the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We thank Dr. Juçara Parra for her general administrative coordination of this program. We are also indebted to LNCC and LICR for their administrative and technical support.

REFERENCES