REVIEW ARTICLE

Cloning and assembly strategies in microbial genome projects

Lionel Frangeul,[1] Karen E. Nelson,[2] Carmen Buchrieser,[1] Antoine Danchin,[3] Philippe Glaser[1] and Frank Kunst[1]

Author for correspondence: Frank Kunst. Tel: +33 1 45 68 88 69. Fax: +33 1 45 68 87 86. e-mail: fkunst@pasteur.fr

[1] Laboratoire de Génomique des Microorganismes Pathogènes, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris Cedex 15, France

[2] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

[3] Unité de Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France

Overview

The knowledge of an entire genome sequence not only provides a wealth of data, but also specific information that can not be obtained by other approaches. Only after the completion of genome projects did it become obvious that many genes had not been identified by classical genetics. These included many genes that do not share similarities with known genes in databases, paralogous genes leading to redundancy of gene function, genes corresponding to potential new drug targets (Arigoni *et al.*, 1998) and inactive genes resulting from reductive evolution (Andersson *et al.*, 1998). Whole-genome sequencing has revealed important information on the organization of genomes. These include the G+C content and its variation within the genome, leading to the identification of DNA regions acquired by horizontal gene transfer, the presence of repeated elements or insertion elements, the discovery of new pathogenicity islands in pathogenic bacteria and the identification of new operons, genome polarity, and the identification of origin(s) of replication. Genome projects also can be starting points for other projects such as metabolic reconstruction (Selkov *et al.*, 1997) and the systematic functional analysis of all of the genes of an organism which are, for example, currently being carried out for *Saccharomyces cerevisiae* (Dujon, 1998) and *Bacillus subtilis* (see the INRA web site; Table 1).

Several projects involving genome sequencing of model organisms, for example *S. cerevisiae* (Goffeau *et al.*, 1997), *Escherichia coli* (Blattner *et al.*, 1997) and *B. subtilis* (Kunst *et al.*, 1997), have been completed using a large amount of previously available information, including the genetic and physical maps of the organisms as well as ordered large-insert libraries of cosmid, lambda, YAC (yeast artificial chromosome) or BAC (bacterial artificial chromosome) clones.

Nowadays, substantial preliminary data are no longer necessary for the initiation of a genome project. However, a prerequisite for a successful project is detailed pre-planning, and some basic knowledge of the bacterium at the beginning of the project is helpful. Features such as genome size, G+C content, the presence of a circular or linear chromosome and the eventual presence of (circular or linear) plasmids should be taken into account. The strategies to be followed should be very carefully considered, since midstream changes are time consuming and costly. In particular, the choice of the cloning strategies is crucial, since the construction of gene banks of poor quality may be unnoticed in the beginning of the project and lead, at a later stage, to assembly difficulties (see below).

The aim of a microbial genome project is to construct, from 500–800 bp sequence reads containing about 1% mistakes, a genome sequence of several megabases with an error rate lower than 1 per 10 000 nucleotides. Because of technical improvements, including the availability of new software, automated sequencers and low computation costs, a microbial genome project can be completed in a small laboratory within two years.

In the first section of this review we will present a summary of the various cloning and sequencing strategies used in genome projects, including random and ordered approaches. In the following sections we will focus on the whole-genome shotgun-sequencing approach.

Table 1. Web sites cited in the text

Organization	Web address (URL)
The Institute for Genomic Research (TIGR)	http://www.tigr.org/
GOLD, University of Illinois*	http://www.ebi.ac.uk/research/cgg.genomes.html
	http://geta.life.uiuc.edu/∼nikos/genomes.html
Invitrogen	http://www.invitrogen.com/
Caltech Genome Research Laboratory	http://www.tree.caltech.edu/
Laboratory of Genomics of Microbial Pathogens (GMP)	http://www.pasteur.fr/recherche/unites/gmp/
Stratagene	http://www.stratagene.com/
University of Oklahoma†	http://www.genome.ou.edu/big_dyes_bact.html
Clontech Laboratories	http://www.clontech.com/
Phrap, University of Washington	http://bozeman.mbt.washington.edu/
INRA	http://locus.jouy.inra.fr

*GOLD, Genome On Line Database.

†Sequencing of bacterial genomic DNA with ABI's big dye terminators.

Genome-sequencing strategies

Two genome-sequencing strategies have frequently been used (Fig. 1). The first is the ordered-clone approach that uses a large-insert library to construct a map of overlapping clones covering the whole genome; selected clones are then sequenced to obtain the whole-genome sequence. The second strategy is direct shotgun sequencing, which does not require preliminary data (such as a map) before the sequencing phase.

Ordered-clone approach

Several methods are used to order clones, including restriction fingerprinting and hybridization mapping. Fingerprinting is a procedure for clone comparison based on the matching of characteristic restriction fragment sets ('fingerprints'). If the fingerprints of two clones are found to share a significant number of restriction fragments, it may be assumed that these two clones overlap and form a contiguous sequenced region, called a contig (Gregory *et al.*, 1997). This method has been successfully applied during the *Caenorhabditis elegans* (Ellis *et al.*, 1986) and *Mycobacterium tuberculosis* projects (Cole *et al.*, 1998). Bioinformatic tools such as FPC, MCD and ATLAS (Gillett *et al.*, 1996; Soderlund *et al.*, 1997) may be of help for the construction and editing of clone maps and contigs, using fragment size data from multiple complete endonuclease digestions of a set of clones.

Hybridization is a simple and rapid method for identifying stretches of homologous DNA, and a large number of clones can be analysed and ordered concurrently. With this method, an ordered cosmid library covering nearly the whole genome (8 Mb) of *Streptomyces coelicolor* has been constructed (Redenbach *et al.*, 1996). Chromosomal DNA was digested with a rare-cutting endonuclease, and the resulting gel-purified DNA fragments used as probes to order the cosmid library. Cosmid sublibraries corresponding to each DNA fragment were thus defined and the clones aligned with the help of labelled riboprobes corresponding to the insert ends. These probes can be easily prepared since the cloning site of the vector is flanked by T3 and T7 promoters adjacent to the insert ends. Management of the data may be facilitated by using software such as Cloneplacer (Singh & Krawetz, 1995).

Random-sequencing approach

Currently, the most widely used strategy for the sequencing of a microbial genome is that of whole-genome shotgun sequencing. A large number of clones, from libraries representative of the whole genome, are sequenced and assembled into contigs. The contigs are then joined together using a variety of methods to obtain the whole genome sequence in a single contig. It should be kept in mind that 90–95% of the sequence, collected during the shotgun phase of the project, is usually assembled into several hundred contigs without any difficulty. However, the obtention of the final 5–10% of the sequence and its assembly into a single contig is an important task that requires careful planning, as is outlined throughout this review.

With the use of the random-sequencing approach, the complete genome sequences of several organisms have been obtained, including *Haemophilus influenzae* (1·83 Mb), *Helicobacter pylori* (1·66 Mb), *Archaeoglobus fulgidus* (2·18 Mb) and *Thermotoga maritima* (1·86 Mb) (Fleischmann *et al.*, 1995; Klenk *et al.*, 1997; Nelson *et al.*, 1999; Tomb *et al.*, 1997; see also the TIGR and University of Illinois web sites; Table 1). This method has now been adopted by many laboratories for the completion of genomes in the size range of 2–6 Mb, and possibly also beyond this size since the utilization of this method has been discussed for the human genome (Green, 1997; Weber & Myers, 1997).

In the following sections of this review we describe the various phases of a whole-genome shotgun-sequencing
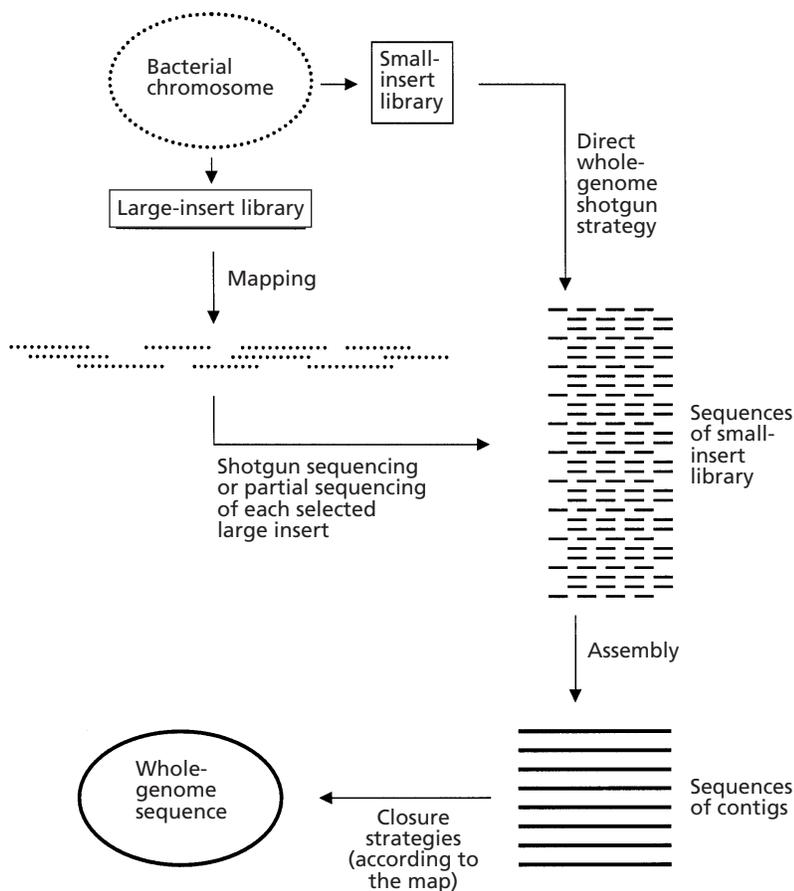
Fig. 1. Strategies used to obtain whole-genome sequences: the ordered-clone approach with the help of a map constructed from a large-insert library and the direct shotgun approach with the utilization of a small-insert library.

project, including the cloning step, the random-sequencing phase, the assembly of sequences into contigs and the closure phase where the contigs are assembled into a single contig spanning the entire genome sequence. The precautions to be taken to ensure completion of a genome-sequencing project and the difficulties encountered during each phase of the project will be discussed.

Library construction

Experience from completed genome projects has demonstated the need for several libraries (Fleischmann *et al.*, 1995). The recommended procedure is to construct a small- and a large-insert library concurrently. The small-insert library is used for an appropriate coverage of the genome and the large-insert library to obtain a 'scaffold' of the genome which is used during the closure phase.

The first precaution to be taken before starting the construction of a library is to ensure that the genome of the chosen organism is stable under the culture conditions used. In bacteria with a short doubling time in rich media (less than 40 min), for example *E. coli*, *B. subtilis* and many others, the region surrounding the origin of replication will be overrepresented four to eight times compared to the replication termination region. To avoid such a potential bias in gene representation in genomic DNA libraries, it may be advisable to extract genomic DNA from a culture in the early stationary phase. However, an argument against this approach is that the cell walls of some stationary-phase bacteria are more difficult to lyse than those of exponentially growing bacteria. Additionally, cloning problems may result from the base composition (A + T-rich regions) or structural features of the DNA (palindromic sequences). Such problems may be solved by generating libraries with very small inserts (200–500 bp) (McMurray *et al.*, 1998), or PCR amplification. Special precautions are recommended for the manipulation and cloning of A + T-rich DNA. These include extraction of DNA under high-salt conditions to prevent DNA from melting at a high temperature, the use of a reduced extension temperature in PCR and the use of transposon-insertion methods for the closure of gaps that are very rich in A + T (Gardner *et al.*, 1998).

Small-insert libraries

It is important to construct a small-insert (1–2 kb) library with a single genomic DNA insert in each recombinant clone. The presence of multiple insertions in a given clone would cause genome-assembly artifacts. To avoid tandem inserts of two independently cloned fragments (co-cloning), Fleischmann *et al.* (1995) developed a procedure based on two steps of size selection.

The *Bst*XI non-palindromic cloning method is another efficient strategy which minimizes co-cloning. This

method has been applied to several genome projects (Smith *et al.*, 1997) and involves the use of special vectors containing a polylinker region with two *Bst*XI sites, for example the ColE1 derivatives pcDNA1.1/ Amp or pcDNA2.1 (Invitrogen; Table 1). After *Bst*XI cleavage, the vector fragment containing non-palindromic TGTG-3′ overhanging ends is purified and ligated to chromosomal DNA that has been fragmented by nebulization (Bodenteich *et al.*, 1994). However, prior to ligation, gel-purified fragments in the range of 1–2 kb have their overhanging ends filled-in by T4 DNA polymerase, then they are ligated to CACA-3′ adaptors and repurified. Under these conditions, a high proportion of the obtained recombinant plasmids contain single DNA inserts because the vector fragment can not recircularize and the probability of adaptor addition to a single blunt-ended DNA fragment is substantially higher than the ligation of two blunt-ended DNA fragments.

Large-insert libraries

High-level expression of genes whose products are toxic to the *E. coli* host (Kurland & Dong, 1996) is a major cause of cloning problems, especially with large-insert libraries (McMurray *et al.*, 1998). This has been shown to be the case for genes encoding membrane proteins (Beck & Bremer, 1980; Beucher & Sparling, 1995; Luo *et al.*, 1997). The problem of overexpression is particularly severe when genes from Gram-positive bacteria are cloned in a Gram-negative host. Gram-positive bacteria use ribosomal binding sequences which are close to the consensus sequence for efficient translation initiation (Farwell *et al.*, 1992; Isono & Isono, 1976). This can be avoided, or at least reduced, by using a low-copy-number vector, in particular when a large-insert library is required to provide a scaffold of the entire bacterial genome.

Large fragments (20–300 kb) are usually cloned in lambda, cosmid, fosmid or BAC vectors. The main advantages of BAC vectors, which have the origin of replication of the *E. coli* F factor (Shizuya *et al.*, 1992), are the potential to clone inserts of up to 300 kb and their stable maintenance. Moreover, BAC vectors have a copy number of one to two copies per cell, reducing the potential for recombination between cloned DNA fragments and avoiding counterselection due to overexpression of genes (Brosch *et al.*, 1998). Their drawback is the somewhat laborious construction of BAC libraries: shearing of DNA must be avoided and *n vitro* manipulations such as restriction digestions are therefore performed in agarose plugs. A convenient BAC vector (pBeloBAC11) has been described by Kim *et al.* (1996) and detailed information on this vector, as well as protocols for BAC library construction, are available from the California Institute of Technology web site (Table 1). Alternative low-copy-number vectors are derivatives of pRK290, which exist at 5–8 copies per chromosomal equivalent in *E. coli*. Such vectors have been shown to maintain stably foreign DNA up to 300 kb (Tao & Zhang, 1998). Fosmid vectors are BAC

vectors that contain cos sites, facilitating their *in vitro* packaging into lambda (Kim *et al.*, 1992). YAC vectors have also been used for the cloning and stable maintenance of bacterial genomic DNA (Azevedo *et al.*, 1993); however, the low yield of purified YAC recombinant DNA and contamination with yeast genomic DNA have limited their utilization.

Other versatile low-copy-number plasmid vectors are pSYX34 (Xu & Fomenkov, 1994) and pHSG575 (Takeshita *et al.*, 1987). These vectors, which are maintained at about 5 copies per cell, are derivatives of pSC101 (Cohen *et al.*, 1973), contain multiple cloning sites and allow for *lacZ* α-complementation. We have found pSYX34 to be an efficient vector for the construction of a large-insert library (6–12 kb) for the DNA of Gram-positive bacteria. In the case of the *Listeria monocytogenes* sequencing project, carried out by a consortium of European laboratories (GMP web site; Table 1), the partial fill-in method was used for the construction of a large-insert library (Korch, 1987). The pSYX34 vector was cleaved with *Sal*I and the ends partially filled-in using Klenow polymerase to create 5′-TC overhangs. Chromosomal DNA from *L. monocytogenes* was partially digested with *Sau*3AI and partially filled-in to create 5′-GA overhangs. After *in vitro* ligation and transformation, 86 % of the clones displayed a white phenotype and about 90 % of these contained inserts in the expected size range of 6–12 kb (our unpublished results). In contrast, when using a high-copy-number plasmid such as pBluescript (Stratagene; Table 1), the percentage of white clones was much lower (9 %), and only 25 % of these contained inserts. These results show the advantage of a low-copy-number vector such as pSYX34 for the construction of large-insert libraries, especially in the case of DNA from Gram-positive bacteria. Cloning difficulties associated with high copy ColE1-type vectors can also be avoided using *E. coli* strains in which the copy number of these plasmids is lowered. The copy number of ColE1 derivatives is strongly reduced in the *pcnB cya* mutant TP611 (Glaser *et al.*, 1993) and is reduced about fourfold and 10-fold in the ABLE C and K strains, respectively, commercially supplied by Stratagene (Table 1).

Assembly phase

During the assembly and closure phases, both computational and experimental results are crucial for obtaining the entire genome sequence. The various software tools used during these steps are listed on the GMP web site (Table 1). It should be kept in mind that the choice of an appropriate bioinformatics package should be made at the beginning of the project, since changing to another package generally leads to a vast amount of additional work.

The assembly phase is composed of three major steps: the conversion of the data from automated sequencers to nucleotide sequences, the utilization of these sequences in the assembly process and the continuous assessment of this assembly process.

Utilization of the data supplied by automated sequencers

A large diversity of programs or packages has been developed for the various stages from gel-image analysis to assembly editing. The analysis of the gel image is generally performed by the software supplied with the sequencer. However, it can also be analysed by other software tools such as Getlanes (Cooper *et al.*, 1996) or GelImager (Giddings *et al.*, 1998), which allow an automatic or semi-automatic separation of low-quality sequences from those of adequate quality. The following step, 'base-calling', is critical for obtaining high-quality sequences suitable for subsequent annotation of the contigs. In the past, for each chromatogram extracted from the gel image, users had to discriminate between high and low quality sequence areas. Recently, the development of programs such as BaseFinder (Giddings *et al.*, 1998) or Phred (Ewing & Green, 1998; Ewing *et al.*, 1998) has led to significant improvements in these procedures. These software tools analyse the traces of each sequence lane automatically and deliver a nucleotide sequence with an accuracy value for each nucleotide. During assembly, the accuracy value associated with each base allows the use of the entire sequence extracted from the traces. The weight of one base in the multi-alignment is proportional to the accuracy value provided by the base-caller.

The appropriate functioning of bioinformatic tools may depend on the type of automated sequencer used. Some base-callers have been created and validated on the chromatograms generated by one type of automated sequencer, leading to a slight bias of the software to the type of apparatus used. For instance, in our hands, the Phred base-caller produces better results with PE-ABI chromatograms compared to the standard chromatogram format originating from sequencers using the single-colour fluorescent technology.

Before the assembly step, cloning vector sequences are removed from the sequences supplied by the base-caller. These vector sequences can be masked automatically by several software tools such as Cross-match which uses the Smith and Waterman algorithm (University of Washington web site; Table 1).

Assembly step

The assembly of more than 10000 sequences into a few contigs, typically required for a bacterial genome project, is based on complex algorithms. The following need to be taken into account: the error rate of each sequence, the use of the generated sequence itself or its complement, and the presence of repeated sequences whose missassembly needs to be avoided. Computer tools to solve these problems were first described by Staden (1979, 1982) and have been intensively developed over the last few years (Kececioglu & Myers, 1995). The most recent algorithms perform pairwise comparisons for all sequences, allowing automatic threshold selection with respect to the decision of whether two sequences overlap or not. Clusters of overlapping sequences are constructed and consensus sequences are deduced from these clusters. The advanced algorithms allow the location of repeated and chimeric sequences. The most widely used assembly software tools are the TIGR-Assembler (Sutton *et al.*, 1995), CAP2 (Huang, 1996), GAP (Bonfield *et al.*, 1995) and Phrap (P. Green, University of Washington, Seattle, USA, unpublished). To take advantage of each software, efforts have been made to allow data exchanges between some of the packages. For example, TIGRAssembler can create an assembly-result file in Phrap format so that it can be viewed with Consed (Gordon *et al.*, 1998). Moreover, tools such as Caftools (Dear *et al.*, 1998) have been created to facilitate such exchanges. In 1994, a comparison of 11 assembly software tools showed important discrepancies between their results (Miller & Powell, 1994). However, to our knowledge, no recent comparisons have been performed on currently available software tools.

Providing there is no cloning bias, the assembled DNA fragments are located around the chromosome according to a Poisson distribution (Lander & Waterman, 1988; Fraser & Fleischmann, 1997). The fraction of the genome which remains unsequenced on both strands, is

$$p_0 = e^{-nw/L}$$

where n is the number of sequenced clones, w is the the mean length of a sequence and L is the length of the genome (in kb). A threefold coverage of the genome ($nw/L = 3$) would therefore produce about 95 % ($p_0 = 0.05$) of the entire genome sequence on the basis of a random distribution. This level of coverage should already be sufficient for the discovery of genes of interest. If the final goal of the project is the completion of a genome sequence, a higher coverage, up to eightfold, is needed for the accomplishment of the shotgun phase. This would require about 15000 reads of 500 bp per megabase of genome length and would produce theoretically more than 99 % of the genome sequence. The number of contigs, which equals the number of unsequenced regions (double-strand gaps), can be calculated as

no. contigs = no. gaps = $ne^{-nw/L}$ (Fraser & Fleischmann, 1997)

For a 4 Mb genome, the predicted number of contigs after a sevenfold coverage is about 50 ($n = 56000$, $L = 4000$ kb, $w = 0.5$ kb). In practice, the number of contigs is generally higher since several criteria are not taken into consideration in this calculation. The assembly software needs an overlap of several nucleotides to link two clones and the representation of some DNA regions may be biased due to cloning difficulties. In this context, an important effort must be dedicated to the assessment of the assembly, as explained below.

Assessment of the assembly

Several software tools such as Consed (Gordon *et al.*, 1998) are dedicated to editing the assembly results. They show the structure of the assembly (the location of all

clones inside the contigs) with direct links to the traces of each clone and allow tags to be added to the sequences. Some of these tags change the accuracy value attributed to the sequences of a clone, and consequently affect the result of the next assembly run.

At low coverage levels of about onefold, intermediate assemblies should be carried out as tests: a large number of repeat sequences may occur in the genome and it is preferable to find solutions to this problem at an early stage of the project. The simplest way to avoid mis-alignment due to short repeated sequences is to obtain sequence reads that span most of the repeats. After assembly, most long repeats are found not to be exact repetitions. In this case, it is important to check contig regions containing several clones with mismatches in high-quality (i.e. high confidence values) sequence areas. If these regions correspond to a false assembly, the assembly software can be re-run with instructions to separate these clones.

Similarly, the locations of the end sequences of each clone have to be verified. In most cases, these sequences are part of the same contig containing the first sequence and the complement of the second sequence. A further requirement is that these two sequences are separated by a distance compatible with the cloned-insert length. Two sequences separated by an inconsistent distance and/or oriented in the same sense indicate a potential misalignment. A visual inspection of the corresponding regions should be performed and the misalignment corrected using tags in the sequences to allow the software to perform a better assembly.

There are no clear bases for deciding when to shift from the random strategy to the closure phase. This decision depends on the following considerations: does the contig number decrease sufficiently in proportion to the number of new sequences added to the assembly; is the mean coverage of each contig sufficient; are the closure strategies expected to be successful at an acceptable cost or would it be more advantageous to delay the shift from the random to the closure phase?

A thorough analysis of the sequences obtained by the assembly software allows for an effective choice of the moment to start the closure phase, which of course can be performed concurrently with the end of the shotgun sequencing phase.

Closure phase

After contig assembly, the first step is to order the contigs and to try to link them together with specific PCR products or cloned inserts that span each gap. The resulting contigs that are still unlinked are extended using methods as described below. Finally, a single contig is obtained. Fig. 2 summarizes the various methods used to identify the potential neighbours of contigs.

During the random phase, both ends of small- and large-insert clones are sequenced. Thus, if the terminal
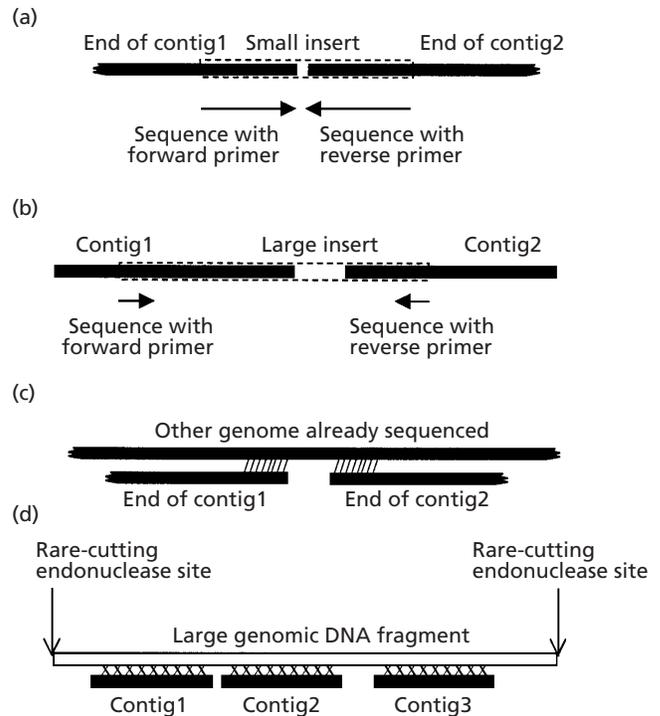

Fig. 2. Methods for the construction of supercontigs. (a) Contigs sharing sequences with a linking small-insert clone. (b) Contigs sharing the end sequences of a linking clone from a large-insert library. (c) Contigs sharing the same operon (or gene) in another entirely sequenced genome. (d) Contigs identified by hybridization to be located on the same large genomic fragment. The symbols used are: cloned insert of the linking clone (rectangle with dotted lines); sequences performed on these clones (arrows); known sequences (black boxes); unknown sequences (white boxes); similarity detected by hybridization (×××××××); similarity detected by BLAST (/////).

sequences of a single clone belong to different contigs, it is highly probable that these two contigs are neighbours (Fig. 2a, b). However, the orientation of the sequences and the distance to the end of the contig should be compatible with the size range of the inserts. These clones are called 'linking clones' and the resulting contigs are 'supercontigs'. The probability that one clone can be used as a linking clone increases with its size. Therefore large-insert libraries are more useful at this stage in the assembly process.

Whatever the coverage of the genome, physical gaps will remain due to unclonable regions. However, the contig order can be deduced from the analysis of coding sequences (CDSs) predicted for each contig. If the comparison between protein databases and the CDSs from contig ends show that the ends of two contigs encode different parts of the same protein, it may be assumed that these contigs are neighbours. In addition, the gene order in certain operons may also be conserved between the genome being sequenced and the sequenced genome of a related organism, hereafter called a control

genome (Fig. 2c). In this way, a control genome may be of help in the completion of the genome under study. For the informatic part of this strategy we have developed a software tool, called GMPTB. For this program, the FASTA database of the contig sequences created by Phrap is used as the input file. The user must choose the length of contig ends to be taken into account for the calculation (e.g. 1000–2000 nt). GMPTB then translates the largest ORF in this terminal region and searches for the counterpart of this protein in the protein database of the control genome with BLASTP (Altschul *et al.*, 1997). If a significantly similar protein is found (>20% identity over the total length of the smaller of the two proteins), GMPTB searches for the location of the gene encoding this protein in the control genome. Finally, all pairs of selected genes from the genome under study whose counterparts are less than 10000 nucleotides apart in the control genome, are transferred to a results file. This file lists all these pairs with their homologous contig ends, which are therefore likely to be neighbours. In the case of *L. monocytogenes*, the protein sequences were systematically compared with those of *B. subtilis*, whose genome has already been completed (Kunst *et al.*, 1997). At least 20% of a sample of 220 *in silico* predicted contig links were shown to be authentic, since these have been confirmed experimentally by the identification of large-insert clones linking these contigs (our unpublished data). With the increasing number of fully sequenced genomes, it is likely that this strategy will be applicable to many other projects.

If computational approaches fail to predict links between the remaining contigs, experiments may be required. Southern-hybridization experiments, using probes corresponding to different contigs, may show that some of these contigs are located on the same large restriction fragment (Fig. 2d). Such fragments can be obtained after digestion of chromosomal DNA with a rare-cutting endonuclease and separated by PFGE (Ramos-Diaz & Ramos, 1998; Umelo & Trust, 1998; Ze-Ze *et al.*, 1998). Other techniques for rapid mapping, including optical mapping and HAPPY mapping may be envisaged (Cai *et al.*, 1998; Piper *et al.*, 1998). These experimental approaches are time-consuming and may be unnecessary in the case of small genomes.

All potential neighbour predictions have to be verified by standard or long-range PCR (Barnes, 1994). For the contigs without identified neighbours, gaps in genomic sequences may be bridged by chromosomal-walking methods. Direct sequencing of the chromosomal DNA template (Heiner *et al.*, 1998; see also the University of Oklahoma web site; Table1) has proved to be a powerful technique, allowing the rapid completion of the *T. maritima* genome project (Nelson *et al.*, 1999). Another method for closing gaps is inverse PCR (standard and long-range; Offringa & van der Lee, 1995). This method allows the extension of contigs by amplification of regions flanking known DNA sequences. Both inverse and long-range PCR have been used extensively for the completion of the *B. subtilis* genome project (Bolotin *et al.*, 1996; Ogasawara *et al.*, 1994; Petit *et al.*, 1998; Rose

& Entian, 1996; Sorokin *et al.*, 1996). An interesting and efficient alternative to inverse PCR is ligation-mediated PCR. This single-sided amplification method involves the ligation of a linker to DNA restriction fragments and subsequent amplification with a pair of primers, one of which recognizes the linker and the other a genomic sequence (Prod'hom *et al.*, 1998). A refinement of the adaptor-ligation method, allowing reproducible amplification without background problems, has been published by Siebert *et al.* (1995) and a GenomeWalker kit based on this method has been commercialised by Clontech (Table 1).

An extension of the PCR methods, combinatorial PCR (multiplex PCR) has also been developed. It involves PCR amplification using a mixture of (up to 32) primers that are located in the vicinity of contig ends. These primers are designed to obtain the simultaneous amplification of PCR products, each primer pair allowing linkage of two contigs. Parallel PCR reactions are carried out in which one primer at a time is left out from the mixture. The absence of a given PCR product in these control reactions will identify the primer left out as being required for the synthesis of a particular product. After identification of each pair of productive primers, the contigs can be ordered and the gaps filled after sequencing of the obtained PCR products (Sorokin *et al.*, 1996).

Whatever the PCR method used, software tools such as Consed or more specific programs such as Primo (Li *et al.*, 1997) or Pride (Haas *et al.*, 1998) automatically choose the primers needed for the closure phase; the primers are chosen according to the usual criteria (selectivity, melting temperature, etc.) and also according to the accuracy values of contig sequences to ascertain that the primer sequences correspond to genome sequences likely to be devoid of errors. Consed can also display the regions that have to be covered by additional sequences. These include regions with low-quality coverage, regions covered only by a single clone and regions sequenced only by dye-primer technology in a single orientation, since this technology leads to an increased frequency of sequence compressions.

Conclusion

Genomics is a new field of biology which has expanded very rapidly in the past few years. Technological advances in automated sequencing and in molecular biology (gene cloning techniques, construction of genome maps, etc.) together with the development of bioinformatics, has ensured the dynamism of this domain of biology. About 20 genome projects have been completed and more than 70 projects are in progress (see the TIGR and the University of Illinois web sites; Table 1). As more genome sequences become available, it is obvious that comparative genomics will play an increasingly important role in the closure phase of a new genome project. In this context, new software tools will be developed to meet the demands of this rapidly moving field of biology.

The completion of a genome sequence represents the end of one phase in the journey of discovery, and important tasks still lie ahead. These include both *in silico* analyses (homology searches) and experimental analyses for the deduction of unknown metabolic pathways, the prediction of operon structures and the identification of regulatory signals. DNA microarray techniques and two-dimensional proteomic analysis will also be used for monitoring gene activity under various environmental conditions. Additionally, the knowledge of a complete genome may identify gene displacements and horizontal transfers, which may help to trace evolutionary networks and to define common ancestors. The complete genome sequence of an organism provides a wealth of information that will provide the basis for a new generation of fundamental and applied biological studies.

Acknowledgements

References

Altschul, S. F., Madden, T. L., Schaffer, 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.

Andersson, S. G. E., Zomorodipour, A., Anderssen, J. O. & 7 other authors (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140.

Arigoni, F., Talabot, F., Peitsch, M. & 7 other authors (1998). A genome-based approach for the identification of essential bacterial genes. *Nat Biotechnol* **16**, 851–856.

Azevedo, V., Alvarez, E., Zumstein, E., Damiani, G., Sgaramella, V., Ehrlich, S. D. & Serror, P. (1993). An ordered collection of *Bacillus subtilis* DNA segments cloned in yeast artificial chromosomes. *Proc Natl Acad Sci USA* **90**, 6047–6051.

Barnes, W. M. (1994). PCR amplification of up to 35 kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc Natl Acad Sci USA* **91**, 2216–2220.

Beck, E. & Bremer, E. (1980). Nucleotide sequence of the gene *ompA* coding the outer membrane protein II of *Escherichia coli* K-12. *Nucleic Acids Res* **8**, 3011–3027.

Beucher, M. & Sparling, P. F. (1995). Cloning, sequencing, and characterization of the gene encoding FrpB, a major iron-regulated, outer membrane protein of *Neisseria gonorrhoeae*. *J Bacteriol* **177**, 2041–2049.

Blattner, F. R., Plunkett, G., III, Bloch, C. A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.

Bodenteich, A., Chissoe, S., Wang, Y. F. & Roe, B. A. (1994). Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. In *Automated DNA Sequencing and Analysis Techniques*. Edited by M. Adams, C. Fields and J. C. Venter. San Diego, CA: Academic Press.

Bolotin, A., Sorokin, A. & Ehrlich, S. D. (1996). Mapping of the 150 kb *spoIIIC–pheA* region of the *Bacillus subtilis* chromosome using long accurate PCR and three yeast artificial chromosomes. *Microbiology* **142**, 3017–3020.

Bonfield, J. K., Smith, K. & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res* **23**, 4992–4999.

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

Cai, W., Jing, J., Irvin, B. & 8 other authors (1998). High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proc Natl Acad Sci USA* **95**, 3390–3395.

Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci USA* **70**, 3240–3244.

Cole, S. T., Brosch R., Parkill J. & 39 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

Cooper, M. L., Maffitt, D. R., Parsons, J. D., Hillier, L. & States, D. J. (1996). Lane tracking software for four-color fluorescence-based electrophoretic gel images. *Genome Res* **6**, 1110–1117.

Dear, S., Durbin, R., Hillier, L., Marth, G., Thierry-Mieg, J. & Mott, R. (1998). Sequence assembly with CAFTOOLS. *Genome Res* **8**, 260–267.

Dujon, B. (1998). European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome. *Electrophoresis* **19**, 617–624.

Ellis, R. E., Sulston, J. E. & Coulson, A. R. (1986). The rDNA of *C. elegans*: sequence and structure. *Nucleic Acids Res* **14**, 2345–2364.

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

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

Farwell, M. A., Roberts, M. W. & Rabinowitz, J. C. (1992). The effect of ribosomal protein S1 from *Escherichia coli* and *Micrococcus luteus* on protein synthesis in vitro by *E. coli* and *Bacillus subtilis*. *Mol Microbiol* **6**, 3375–3383.

Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fraser, C. M. & Fleischmann, R. D. (1997). Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**, 1207–1216.

Gardner, M. J., Tettelin, H., Carucci, D. J. & 22 other authors (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132.

Giddings, M. C., Severin, J., Westphall, M., Wu, J. & Smith, L. M. (1998). A software system for data analysis in automated DNA sequencing. *Genome Res* **8**, 644–665.

Gillett, W., Hanks, L., Wong, G. K. S., Yu, J., Lim, R. & Olson, M. V. (1996). Assembly of high-resolution restriction maps based on multiple complete digests of a redundant set of overlapping clones. *Genomics* **33**, 389–408.

Glaser, P., Kunst, F., Arnaud, M. & 14 other authors (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325 degrees to 333 degrees. *Mol Microbiol* **10**, 371–384.

Goffeau, A. & others (1997). The yeast genome directory. *Nature* **387** (suppl.), 5–105.

Gordon, D., Abajian, C. & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195–202.

Green, P. (1997). Against a whole-genome shotgun. *Genome Res* **7**, 410–417.

Gregory, S., Soderlund, C. & Coulson, A. (1997). Contig assembly by fingerprinting. In *Genome Mapping: A Practical Approach*, pp. 227–254. Edited by P. Dear. Oxford: Oxford University Press.

Haas, S., Vingron, M., Poustka, A. & Wiemann, S. (1998). Primer design for large scale sequencing. *Nucleic Acids Res* **26**, 3006–3012.

Heiner, C. R., Hunkapiller, K. L., Chen, S. M., Glass, J. I. & Chen, E. Y. (1998). Sequencing multimegabase-template DNA with BigDye terminator chemistry. *PCR Methods Appl* **8**, 557–561.

Huang, X. (1996). An improved sequence assembly program. *Genomics* **33**, 21–31.

Isono, K. & Isono, S. (1976). Lack of ribosomal protein S1 in *Bacillus stearothermophilus*. *Proc Natl Acad Sci USA* **73**, 767–770.

Kececioglu, J. & Myers, E. (1995). Combinatorial algorithms for DNA sequence assembly. *Algorithmica* **13**, 7–51.

Kim, U. J., Shizuya, H., de Jong, P. J., Birren, B. & Simon, M. I. (1992). Stable propagation of cosmid-sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20**, 1083–1085.

Kim, U. J., Birren, B. W., Slepak, T., Mancino, V., Boysen, C., Kang, H. L., Simon, M. I. & Shizuya, H. (1996). Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34**, 213–218.

Klenk, H. P., Clayton, R. A., Tomb, J. F. & 48 other authors (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370.

Korch, C. (1987). Cross index for improving cloning selectivity by partially filling in 5′-extensions of DNA produced by type II restriction endonucleases. *Nucleic Acids Res* **15**, 3199–3220.

Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Kurland, C. G. & Dong, H. (1996). Bacterial growth inhibition by overproduction of protein. *Mol Microbiol* **21**, 1–4.

Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239.

Li, P., Kupfer, K. C., Davies, C. J., Burbee, D., Evans, G. A. & Garner, H. R. (1997). PRIMO: A primer design program that applies base quality statistics for automated large-scale DNA sequencing. *Genomics* **40**, 476–485.

Luo, Y., Glisson, J. R., Jackwood, M. W., Hancock, R. E., Bains, M., Cheng, I. H. & Wang, C. (1997). Cloning and characterization of the major outer membrane protein gene (*ompH*) of *Pasteurella multocida* X-73. *J Bacteriol* **179**, 7856–7864.

McMurray, A. A., Sulston, J. E. & Quail, M. A. (1998). Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res* **8**, 562–566.

Miller, M. J. & Powell, J. I. (1994). A quantitative comparison of DNA sequence assembly programs. *J Comput Biol* **1**, 257–269.

Nelson, K. E., Clayton, R. A., Gill, S. R. & 24 other authors (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329.

Offringa, R. & van der Lee, F. (1995). Isolation and characterization of plant genomic DNA sequences via (inverse) PCR amplification. *Methods Mol Biol* **49**, 181–195.

Ogasawara, N., Nakai, S. & Yoshikawa, H. (1994). Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Res* **1**, 1–14.

Petit, M. A., Dervyn, E., Rose, M., Entian, K. D., McGovern, S., Ehrlich, S. D. & Bruand, C. (1998). PcrA is an essential DNA helicase of *B. subtilis* fulfilling functions both in repair and rolling-circle replication. *Mol Microbiol* **29**, 261–273.

Piper, M. B., Bankier, A. T. & Dear, P. H. (1998). A HAPPY map of *Cryptosporidium parvum*. *Genome Res* **8**, 1299–1307.

Prod'hom, G., Lagier, B., Pelicic, V., Hance, A. J., Gicquel, B. & Guilhot, C. (1998). A reliable amplification technique for the characterization of genomic DNA sequences flanking insertion sequences. *FEMS Microbiol Lett* **158**, 75–81.

Ramos-Diaz, M. A. & Ramos, J. L. (1998). Combined physical and genetic map of *Pseudomonas putida* KT2440 chromosome. *J Bacteriol* **180**, 6352–6363.

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

Rose, M. & Entian, K. D. (1996). New genes in the 170 degrees region of the *Bacillus subtilis* genome encode DNA gyrase subunits, a thioredoxin, a xylanase and an amino acid transporter. *Microbiology* **142**, 3097–3101.

Selkov, E., Maltsev, N., Olsen, G. J., Overbeek, R. & Whitman, W. B. (1997). A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* **197**, GC11–26.

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

Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. & Lukyanov, S. A. (1995). An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* **23**, 1087–1088.

Singh, G. B. & Krawetz, S. A. (1995). CLONEPLACER: a software tool for simulating contig formation for ordered shotgun sequencing. *Genomics* **25**, 555–558.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C. & 34 other authors (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: *functional analysis and comparative genomics*. *J Bacteriol* **179**, 7135–7155.

Soderlund, C., Longden, I. & Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13**, 523–535.

Sorokin, A., Lapidus, A., Capuano, V., Galleron, N., Pujic, P. & Ehrlich, S. D. (1996). A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. *Genome Res* **6**, 448–453.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601–2610.

Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Res* **10**, 4731–4751.

Sutton, G., White, O., Adams, M. & Kerlavage, A. (1995). TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci Technol* **1**, 9–19.

Takeshita, S., Sato, M., Toba, M., Masahashi, W. & Hashimoto-Gotoh, T. (1987). High-copy-number and low-copy-number plasmid vectors for *lacZ* alpha-complementation and chloramphenicol- or kanamycin-resistance selection. *Gene* **61**, 63–74.

Tao, Q. & Zhang, H.-B. (1998). Cloning and stable maintenance of

DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. *Nucleic Acids Res* **26**, 4901–4909.

Tomb, J. F., White, O., Kerlavage, A. R. & 39 other authors (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547.

Umelo, E. & Trust, T. J. (1998). Physical map of the chromosome of *Aeromonas salmonicida* and genome comparisons between *Aeromonas* strains. *Microbiology* **144**, 2141–2149.

Weber, J. L. & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Res* **7**, 401–409.

Xu, S. Y. & Fomenkov, A. (1994). Construction of pSC101 derivatives with Cam^r and Tet^r for selection or LacZ′ for blue/white screening. *Biotechniques* **17**, 57.

Ze-Ze, L., Tenreiro, R., Brito, L., Santos, M. A. & Paveia, H. (1998). Physical map of the genome of *Oenococcus oeni* PSU-1 and localization of genetic markers. *Microbiology* **144**, 1145–1156.