Random PCR-Based Genome Sequencing: A Non-Divide-and-Conquer Strategy

Koichi NISHIGAKI,* Koichiro AKASAKA, and Takashi HASEGAWA

Department of Functional Materials Science, Saitama University, Urawa, Saitama 338-8570, Japan

(Received 18 November 1999)

Abstract

We propose a genome sequencing strategy, which is neither divide-and-conquer (clone by clone) nor the shotgun approach. Random PCR-based and PCR relay sequencing constitute the basis of this novel strategy. Most of the genome is sequenced by the former process that requires only a set of non-specific primers and a template DNA. Random PCR-based sequencing reduces redundancy in sequencing by exploiting known sequence information. The number of primers required for random PCR was significantly diminished by using a combination of primers. The former process can be partially replaced by the shotgun method, if necessary. The gap-filling process can be effectively performed by way of PCR relay. The feasibility of this strategy was demonstrated using the Escherichia coli genome. This strategy enhances the global effort towards genome sequencing by being available through the Internet and by allowing the use of preexisting sequence data.

Key words: genome sequencing; strategy; random PCR; PCR-relay; shotgun

1. Introduction

Genome sequencing projects are being accelerated. Momentous results have been achieved from studies of bacterial and yeast genomes.1-3 Human genome projects in the USA and elsewhere seem to be making great advances.4,5 However, new strategies may yet be proposed.6-12 Here, we propose a new strategy that may not only save a great deal of effort and cost, but also enhance our understanding of the genome. Most strategies so far applied to high-volume genome sequencing have been based on the notion of 'divide and conquer,' which fundamentally requires cloning and mapping. Though establishing a library of clones may offer advantages to subsequent research such as that of gene expression, it is not indispensable for genome sequencing especially when the equivalent effect might be obtained by PCR or by understanding the entire sequence of a genome. If an alternative were available, the laborious and time consuming steps of sequencing could be omitted. Whole shotgun approaches have been advocated for viral and bacterial genomes of around 2 Mb or less.13-15 However, this approach may not be applicable to the human genome.8,10,16 Here, we present a novel non-divide-and-conquer strategy for genome sequencing.

1.1. Strategy

Random PCR-based Genome Sequencing (RGS), is a reinforced derivative of our original strategy, ROSOR (Rapid Oligonucleotide Synthesis and Oligonucleotide Relay), of which the essence is briefly shown in Fig. 1a.17 ROSOR consists of three main steps: Setting multiple starting sites (anchoring sites) on genomic DNA to be sequenced by random PCR,16,19 successive sequencing using DNA fragments (PCR-relay) prepared by single specific primer PCR20 and rapid oligonucleotide synthesis. RGS is far simpler than ROSOR, since RGS requires considerably fewer oligonucleotide primers and it complements conventional oligonucleotide synthesis. The most important step in RGS (Fig. 1b), is sequencing based on random PCR. Sequence data obtained in this manner will accumulate and cover most of the genome to the extent of $\alpha$ (0 ≤ $\alpha$ ≤ 1). In RGS, the remainder of the genome sequence, 1-$\alpha$, can be determined by PCR relay (Fig. 1c). Preparing the vast number of oligonucleotides required for a complete experiment is laborious, time consuming and expensive (such as $10^6$ oligonucleotides [if $(1-\alpha) = 0.1$] and $10^5$ [if $(1-\alpha) = 0.01$]). Although still somewhat costly, this approach is now feasible.

In this paper, we tentatively apply random PCR-based genome sequencing (RGS) to the E. coli genome and discuss its advantage over conventional strategies.
Figure 1. Schematic representation of RGS. (a) Strategic difference between ROSOR and RGS. Spheres indicate regions sequenced via random PCR and arrows, sequenced by PCR relay. ROSOR (Rapid Oligonucleotide Synthesis and Oligonucleotide Relay) is a strategy by which an entire genome can be sequenced in three steps namely, setting-up anchoring sites by random PCR, consecutive sequencing by PCR-relay and rapid synthesis of oligonucleotides. RGS (Random PCR-based Genome Sequencing) consists of sequencing shotgun clones or random PCR products and PCR-relay. (b) The entire RGS process. Each stage is shown by categorization into a random, semi-random or directed phase, although these three phases can run in parallel. The term ‘semi-random’ arises from the fact that the DNAs amplified by random PCR are specifically determined by the applied oligonucleotides (c) PCR relay processes. 1) Genome DNAs are partially digested by restriction enzymes such as Sau3AI. Long hatched box represents sequence determined via shotgun or random PCR. 2) Each DNA fragment is sandwiched by short DNA fragments (oligo-cassettes), that makes a library of X-digested genome DNA fragments (X = Sau3AI or other). 3) and 4) PCR is performed using two species of primers: One primer (upstream) is specific to the site already sequenced by random PCR prior to PCR-relay and is also biotinylated at its 3'-end and the other primer (downstream), specific to the oligo-cassette (short hatched boxes). After separating PCR products by biotin-avidin binding, the resulting products are a series of X-partially digested DNA fragments of which one end is identical but the other is different. These are then directly used as templates for dideoxy terminal sequencing. (To read longer sequences, a longer fragment was isolated from a gel and cloned to make a sequencing template. Although cassette-to-cassette PCR products are abundant, they can be removed because they have no biotin tag.) 5) The primer that binds to the downstream-side adjacent to the upstream-primer binding site was used to select the correct PCR products. 6) Part of the sequence furthest downstream out of that immediately sequenced was selected and used as the sequence for the next upstream primer. We returned to stage 3 and repeated the process. See Materials and Methods for details.
erable caution was taken to avoid DNA contamination. Once PCR mixtures containing all the components except the specific template DNA were UV-irradiated prior to adding the template DNA. The digestion buffer was prepared as follows: 2.3. PCR relay

Random PCR proceeded as described. Briefly, a reaction mixture of 100 μl contained 200 μM dNTP (N = G, A, T, C), 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 2.5 mM MgCl2, 0.1% Triton X-100, 0.2 mg/ml gelatin, 2.5 mM MgCl2, 0.4 mM dNTP (N = G, A, T, C), and 0.04 units/μl Taq DNA polymerase, 0.1 μM primer (each) and 10^-17 M digested E. coli genomic DNA (4/10000 equivalent amount of the above prepared) under PCR conditions: preheating (90°C, 30 sec), 35 × [denaturation (90°C, 30 sec) – annealing (54°C, 1 min) – extension (68°C, 2 min)] and final extension (70°C, 3 min). Our original protocol subjected the PCR products to a selection process based on a specific feature such as avidin-biotin binding. (For this purpose, some specific primers were 5′-biotinylated). However, this step was unnecessary in the current study because the amount of interfering products was quite small (Fig. 3b). One or two major bands resolved by gel electrophoresis of the PCR products were selected, subjected to repeat PCR and sequenced.

2.4. Sequence analysis
DNAs generated by random PCR and PCR relay as described above were cloned in principle through M13 vectors and manually sequenced by conventional dideoxy chain termination (autoradiography) using the BcaBEST dideoxy sequencing kit (TaKaRa, Kyoto). We used Bacillus cardoenearis YT-G DNA polymerase that has been genetically engineered to lose its 5′-to-3′ exonuclease activity. About 20% of the fragments were custom sequenced using a DNA sequencer 377 (Applied Biosystems, Inc.). DNA sequences obtained by either random PCR or PCR relay were analyzed using the BLAST program under the Internet service (DDBJ) supported by the National Institute of Genetics (Mishima, Japan), where the entire E. coli genome sequence has been available since 1997.

3. Results
3.1. Random PCR
The reproducibility of random PCR has already been established. We therefore demonstrate here that PCR products are genuinely derived from a megabase-sized template DNA (E. coli genome DNA) and that a combination of oligonucleotides serve this purpose.

All 10 sets of primers used for this experiment (single primers, S1–S4; double primers, D1–D6) generated DNA fragments that were sequenced and assigned to a portion of the E. coli genome (Fig. 2a). We identified partially complementary primings containing bulges and mismatches, which are consequences of random PCR.
K. Nishigaki, K. Akasaka, and T. Hasegawa [Vol. 7, 22

2.2. PCR relay

The protocol for PCR relay is shown in Fig. 1c. The basic notion of PCR relay has been elucidated by Shyamala and Ames,14 Rosenthal and Jones24 and Collasius et al.,25 although only one or two contiguous steps were described in these papers. The present study demonstrates that this method is applicable to multi-step walking against a megabase DNA. Figure 3 shows that all four groups of contiguous DNA fragments (17 species), arbitrarily selected from the E. coli genome, were in principle generated by single specific primer PCR. Some of these were further confirmed by sequencing (thick lines in Fig. 3a, and 3c–e), although minor non-authentic DNA fragments were occasionally observed (see Fig. 3b). Since the tested regions were arbitrarily chosen, these data show that PCR relay is generally applicable even to such a large genome as that of E. coli. These results are promising, considering that only a single library of Sau3AI restriction fragments was examined. Thus, the finishing problem of gap-filling can be basically solved by PCR relay without depending on established clones.

4. Discussion

When attempting to sequence over 90% of the total genome DNA solely by random PCR, this strategy resembles that of whole genome shotgun.7,13,14 Since the directed phase is more costly than the random or the semirandom process, genomic DNA should be sequenced by the latter process as much as possible. This problem has been addressed26 and we found that over 90% of sequencing should be random, at least under current conditions (since they depend on the cost ratio of the random against the directed process). The divide-and-conquer strategy is associated with problems of making contigs and maps, as well as maintaining and distributing a huge number of clones. RGS as well as the shotgun strategy is free of these disadvantages. In addition, RGS has the other benefits as described below, which are not features of the shotgun approach.

The key advantage of RGS is that the oligonucleotides selected for random PCR can reproducibly extract a specific DNA fragment and that the fragment can be changed simply by altering the primers. This means 5 × 10^7 can be attained by using only 10^4 types of oligonucleotides (n types of oligonucleotides generate hetero-primer pairs of nC2 (= n(n - 1)/2)). Since the number of random PCR products depends on the experimental temperature, multiple products can be generated from a pair of primers by lowering the annealing temperature. This fact is also favorable for reducing the number of oligonucleotides used for RGS. The methodological advantage of RGS over the shotgun approach is discussed below.
Figure 3. PCR relay. (a) Region of *E. coli* genome DNA where PCR relays were performed (*Case 1*). Portion of genomic DNA is shown at the top with Sau3AI restriction sites (downward arrows). Positions of upstream primers (defined in Fig. 1c) are shown above with rightward arrows. DNA fragments that were recovered by PCR, sequenced and contiguously aligned are shown as a bold line with the name of the applied upstream primer (P3, P7-P10). Thin lines attached on the right with a filled box (a-q) denote the PCR products resolved by gel electrophoresis, shown in Fig. 2c. Undetectable bands are shown by 'x'. Filled boxes represent the cassette sequence (see Materials and Methods). (b) Single specific primer PCR products (lanes 2-4). Lane 1, size markers. Primers are P3 (lane 2), P8 (lane 3), P9 (lane 4) and P10 (lane 5). (Results for P7, were obtained from a different gel and are not shown.) Sizes of the assigned fragments are: a(134), b(148), c(298), z under P3(320), e(469), z(525), x under P8(279), j(286), k(418), l(73), m(380), n(242), o(359); p(133), q(292), where nomenclature is as described above. Bands, e, h, k and o were sequenced and confirmed. (c)/(d)/(e); *Case 2*/Case 3*/Case 4* (PCR relay applied to each region of *E. coli* genome DNA, respectively). The definitions used to make these figures (Fig. 3c-e) are the same as those in Fig. 3a except half arrowheads mean that the predicted PCR fragment spanning from the left (or right) end of the bar to the tip of the arrow head, was (solid line) or was not detected (dotted line) by gel electrophoresis. Regions represented by a thick line were sequenced and confirmed. In Fig. 3d, PCR relay proceeded in the reverse-direction. Anchoring sites (i.e., starting sites for PCR relay) are shown as empty boxes. D4-D6 are as described in Fig. 2. *E. coli* genome coordinates for r1 to r26 are: 1499035, 1499215, 1499298, 1499685, 1500045, 1500144, 1500281, 1500585, 3219155, 3219165, 3219234, 3219392, 3219576, 3219614, 3219713, 3219767, 3219824, 3220023, 3220305, 4475771, 4475795, 4476143, 4476842, 4477508 and 4477619. All four DNA tests showed that PCR relay actually works.
that DNA fragments can be specified and catalogued depending on the PCR primers and the gel-mobility of the fragments. The important technical advances are that any DNA fragment obtained by random PCR can be examined and the entire genome can be covered using relatively few oligonucleotides (order of $10^3$ for human genome DNA). In addition, a primer for random PCR can be designed that has a lower probability of generating PCR products derived from the sequenced regions, since random PCR products can be predicted, and provided with the sequences of a template and primers.19 In other words, sequencing redundancy can be reduced since the obtained information can help to select the most promising primers (see Appendix). Evidently, this feature is more valuable at the later stage of genome sequencing when the probability of hitting an unknown DNA is much reduced. [We recently showed that the prediction of random PCR products can be very accurate even when a large genome such as that of E. coli is sequenced (data to be published).] Of practical importance is that RGS needs only a small amount of genomic DNA owing to PCR amplification and does not require the maintenance and distribution of clones. Random PCR has the notable advantage of readily allowing world-wide collaboration since it does not require either preparative steps or initial investment. Worldwide results can be collected and united as a database that holds ‘already sequenced regions’ and ‘primers used for random PCR.’ This type of database can be built upon sequence data obtained by various strategies without conflict or waste assuming the cell lines used are the same. This is one of the most important features of our strategy in terms of large scale, multinational collaboration. A center that distributes information (on correct primer sequences) and collects and processes the sequence information obtained is not difficult to establish since several sites are already performing an equivalent function. Each unit then, from a giant sequencing center to an individual scientist, can contribute to the project depending upon individual resources and capabilities.

The role of oligonucleotides in this study is crucial not only because random and specific types of PCR depends on oligonucleotides, but also because they serve the function of indexing entire genome sequences. Each DNA sequence determined via random PCR can be associated with the applied primer sequence and inversely, the same DNA fragment will be reproduced by random PCR using the primer and the template, which will be helpful for confirming sequences at the finishing stage. Random PCR-based genome sequencing is equivalent to separating whole genomic DNA (N bp) into fragments (f bp in average) that can be totally retrieved (or catalogued) by a set of primers (the number of elements; $\sqrt{2rN/f}$). This is because the combined use of $n$ kinds of primers generates $n(n - 1)/2$ species of DNA fragments, whereas separating by conventional PCR requires $2rN/f$ primers, where $r$ denotes redundancy in sequencing (= $r$ folds of genome coverage). Ideally, human or bacterial genomes can be fragmented by a set of primers consisting of only $10^4$ or $10^5$ elements, respectively.

Steady advances in oligonucleotide synthesis technology allow the preparation of hundreds of oligonucleotides per day with the parallel operation of current DNA synthesizers. The synthesis rate is sufficiently high and fully compatible with a 3 year project requiring $10^8$ species of oligonucleotides. (We devised a rapid and cost-effective way of preparing numerous species of oligonucleotides based on a tetramer library and enzymatic synthesis27). In addition, technological progress such as stretching a read in sequencing help to conserve primers. As a result, the oligonucleotide-dependent sequencing strategy (RGS) has become very potent.

Several problems are often encountered during human genome sequencing such as repeat sequences, polymorphism, introduction of artifact mutations (substitutions, deletions, insertions and rearrangements) and read errors in sequencing. These have recently been addressed in terms of whole genome shotgun.7,8 All strategies including non-divide-and-conquer types cannot be totally free of these disadvantages. Regardless of the approach, tandem repeats require special measures to cope with them, and such measures have not yet generally been devised.10 The assertion that the availability of more clones and subclones will be of considerable help in solving the tandem repeat problem seems plausible, but is groundless. We nevertheless believe that RGS is applicable to human genome sequencing, although much remains to be resolved. We already demonstrated that genome size does not impose particular problems upon genome sequencing from a stochastic viewpoint insofar as the finishing problem can be treated by a directed, rather than a stochastic process.26

Both clone-by-clone and PCR-based technologies have specific weaknesses. The former cannot eliminate mutations once introduced, whereas the latter is rather error-prone. Although the sequence quality of individual sequencing is not so high due to the involvement of random and conventional PCR processes in RGS, this strategy can eventually produce a high quality output of the whole genome sequence owing to high redundancy (3-7 fold) in the coverage (to estimate this value, see our independent paper28). Since PCR relay can ensure in principle, over 3-fold coverage at all sites of the whole genome, each site could be controlled below the error rate of $10^{-4}$, which is an official target value.28 (Incidentally, if two independent sequencing trials generate the same results, then the final error probability becomes close to the product of each individual trial.) Therefore, sequencing redundancy, which is inevitably rather high in RGS, is helpful for correcting PCR-derived mutations. Eventually, both shotgun and random PCR approaches will take more advantage than disadvantage of redundancy as long as it is
Although the cost of each strategy remains difficult to estimate, we tentatively compared the cost of RGS with that of the clone-by-clone approach. Although several unknown factors make estimation difficult, the costs of these approaches do not significantly differ (data not shown), meaning that neither can be ruled out from the perspective of cost. However, RGS can reduce the redundancy effect in whole genome sequencing and can minimize the finishing problems and therefore must have the potential to be less costly than the shotgun and clone-by-clone approaches.

Recently, a clone-dependent strategy, limited seeding [BAC-based shotgun strategy reinforced by sequence-tagged connectors (STC) useful for generating contigs] was proposed to improve the conventional clone-by-clone strategy. In a sense, it is similar to our strategy in that it exploits the random process as much as possible. However, there is a great difference between limited seeding and RGS that the latter does not construct clone libraries. This is the critical difference between the two strategies and the fact that RGS does not depend on clone libraries, makes this strategy both feasible and attractive for world-wide cooperation.

Acknowledgements: This research was partly supported by Grants-in-Aid (#06454661 and #09272203) from the Ministry of Education, Science, Sports and Culture of Japan. One of the authors (KN) is grateful to Dr. Sydney Brenner for intellectual support and encouragement.

Appendix

Knowledge of part of the genome sequence can help increase the probability of hitting undetermined sequences in random PCR.

The deviation of the probability of hitting a target sequence for random PCR (p) from that for a shotgun process (p_s) can be expressed as follows:

\[
\Delta p = p - p_s
\]  

(a1)

For convenience, we introduced the parameter named 'hit confidence (\(C_{hit}\))' as follows:

\[
C_{hit} = (p - p_s)/(1 - p_s) = \Delta p/(1 - p_s) (0 \leq C_{hit} \leq 1) \quad (a2)
\]

This parameter indicates that in a random process (p = p_s) there is no confidence (\(C_{hit} = 0\)) of hitting a target sequence whereas there is full confidence (\(C_{hit} = 1\)) when p = 1. As shown in Fig. A1, a theoretical consideration about \(C_{hit}\) can be made. If primer-binding (binding two facing primers to the appropriate sites) occurs in proportion to the size of the region sequenced or non-sequenced and directly leads to obtaining PCR fragments, the probability of obtaining an unsequenced DNA fragment can be expressed by a shotgun method, p_s, and the probability by RGS, p, as follows:

\[
p_s = (p_2 + (1 - \alpha)p_3)/(p_1 + p_2 + p_3 + p_4)
\]  

(a3)

\[
p = p_2/(p_2 + p_4)
\]  

(a4)

where \(p_1, p_2, p_3, \) and \(p_4\) are the probabilities assigned for each binding type shown in Fig. A1. Based on binomial distribution, each probability can be calculated as follows:

\[
p_1 = (1 - \delta)\alpha^n; \quad p_2 = (1 - \delta)(1 - \alpha)^n; \quad p_3 = (1 - \delta)(1 - \alpha^n - (1 - \alpha)^n); \quad p_4 = \delta
\]  

(a5)

Following the definition of Eq. (a2), we obtain

\[
C_{hit} = (p_1p_2 - p_3p_4 + \alpha(p_2 + p_4)p_3)/(p_2 + p_4)(1 - p_2 + (\alpha - 1)p_4)
\]  

(a6)

Since the value of \(\delta\), the probability of obtaining no products (Type 4 in Fig. A1), is operational and dependent upon the experimental temperature, this value can be reduced to levels such as \(10^{-4}\) by lowering the annealing temperature. For convenience, if \(\delta = 10^{-6} ; \alpha = 0.9; n = 5\); then, \(C_{hit} = 0.91\) and if \(\delta = 10^{-4} ; \alpha = 0.9; n = 5\), then \(C_{hit} = 0.05\). These results indicate that the probability of obtaining a non-sequenced DNA fragment in RGS can be increased, which is useful for reducing redundancy. Furthermore, by using the size-information of PCR products and selecting appropriately sized fragments for sequencing, the value of \(C_{hit}\) (hit confidence) can be experimentally elevated.

References

15. Fraser, C. M. and Fleischmann, R. D. 1997, Strategies for whole microbial genome sequencing and analysis, Electrophoresis, 18, 1207-1216.