Whole genome sequence-enabled prediction of sequences performed for random PCR products of Escherichia coli

Koichi Nishigaki*, Ayumu Saito, Hasegawa Takashi and Mohammed Naimuddin

Department of Functional Materials Science, Saitama University, 255 Shimo-Okubo, Urawa, Saitama 338-8570, Japan

Received January 24, 2000; Revised and Accepted March 7, 2000

ABSTRACT

The sequence of an unknown PCR product generated by random (and conventional) PCR could be determined without sequencing when it is provided with the template DNA sequence. Theoretically, this was based on formerly established ideas which assert that the amount of random PCR product mainly depends on the stability of the primer-binding structures and that the dynamic solution structure of DNA is essentially governed by the Watson–Crick base pairing. However, it has not been clear whether this holds true for larger genomes of mega- to gigabase size, beside the λ phage genome (of 50 kb) used previously, nor has it been ascertained to uniquely specify the sequence of a random PCR product. Here, we jointly use two computer programs together with experimental data from Genome Profiling (i.e. TGGE analysis of random PCR products). The first procedure carried out by a newly remodeled computer program (PCRAna-A1) was shown to be competent to calculate a set of random PCR products from Escherichia coli genome DNA (4.7 Mb). The other procedure performed with another program (Poland-H) played a critical role in determining the final candidate sequence by theoretically offering the initial melting temperature and the melting pattern of unspecified candidate sequences. The success attained here not only proved our method to be useful for sequence prediction but also confirmed the above-mentioned ideas as rational. We believe that this is the first case to computer-utilize a genome sequence as a whole.

INTRODUCTION

One often encounters ‘non-specific’ PCR products in conventional PCR. As PCR technologies become popular, a knowledge of the solution structure dynamics of DNA/RNA has become essential. However, it is not easy to depict the whole image of its dynamic solution structures including less stable ones, especially in the case of large size DNA/RNA. Therefore, it is natural that hitherto studies in this field should have been limited to dealing with rather small sizes with stable structures (1). This has discouraged us in building practical theories as to the solution structure of large DNAs, which would have been useful to understand the PCR phenomenon.

We reported that general PCR products could be predicted based on the postulate that solution structure dynamics were mainly governed by the Watson–Crick base pairing interaction (2). The rightness of this was demonstrated with λ phage DNA (48 502 bp), one of the largest within the whole genome DNAs of known sequence in those days. Though the experiment was rather limited, the conclusion was clear and useful. In short, it clarified the following four important points: (i) the amount of random PCR products can be set to be proportional to the sum of the Gibbs free energy changes of the initial primer-binding structures; (ii) the solution structures of single-stranded DNAs approximate to those governed by only Watson–Crick base pairings, neglecting the other unusual structures such as Hoogsteen types, quadruplexes, pseudoknots and others (3–5); (iii) the thermodynamic parameters available in those days were already sufficiently effective for that purpose; (iv) most of the inhibitory effects for primer binding were provided by a generation of competitive secondary structures which involved the primer-binding sites.

However, there seems to have been a lot of luck involved in the success of the predictions done then due to the limited amount of experimental data together with the small genome size used. Besides, it was not clear whether the same algorithm was also effective for far larger genomes. Since then, rapid advances in genome science have made us all able to use a whole genome sequence of megabase size or larger. Hence, we tried to remodel the former prediction program (PCRA) to make it applicable to even gigabase sized genomes and to thus confirm the former conclusions. In addition, we reinforced the prediction system so as to be able to tell the sequence of general PCR products without sequencing. We used Escherichia coli genome DNA (4.7 Mb) for this purpose.

MATERIALS AND METHODS

Random PCR

PCR operated at a lower annealing temperature like 30°C generates a lot of unidentified DNA fragments, mostly originating from incomplete hybridization of primers to template DNAs.
In the experiments performed here, we chose a temperature method for analyzing the polymorphism of various organisms. This is one of the very useful aspects of this. Based on their thermodynamic stabilities, primer bindings generate a specific, by no means non-specific, DNA fragment inherent to the couple of binding structures. Although such unexpected products are usually called ‘non-specific’, DNA fragment recovery is proportional to the amount of the initial primer-binding structures, which is ideally proportional to the binding constant ($K_b$) in this case (where the concentrations of all competitive primer-binding sites are equal). Since random PCR products are generated as a result of product of two primer-binding events, their amount is proportional to $K_{αβ} = \exp(-\Delta G_{αβ}/kT)\cdot \exp(-1/kT)\cdot (\Delta G_α + \Delta G_β)$, where $K_{αβ}$, $\Delta G_α$, and $\Delta G_β$ are binding constants and $\Delta G$ for primer-binding structures $\alpha$ and $\beta$, respectively.

(Fig. 1). Since the low annealing temperature permits relatively unstable binding structures to be formed, it becomes possible to begin a primer extension reaction with such primer-binding structures. Therefore, as shown in Figure 1, a couple of facing structures, which are extremely diverse. The process of searching for adequate structures which conform to the constraints imposed have been optimized: geometrically possible, abstract binding structures were first calculated out as ‘skeletons’ and then each partial sequence of the template was checked regarding whether it could match with one of the so-called skeletons as if looking for Cinderella by checking whether one has a foot to fit into a particular shoe (skeleton). This algorithm made the computation of random PCR products faster by more than an order of magnitude in cases of the calculation for a primer of small size (15 nt). In addition, computation time, in principle, is proportional to the size of the genome sequence, which means it is applicable to even the human genome.

**Temperature gradient gel electrophoresis (TGGE)**

DNA samples (~10 pmol) added together with an internal reference (204 bp) were charged horizontally on the top edge of a slab gel [4% polyacrylamide (180 × 180 × 1 mm$^3$) containing 8 M urea]. In a brief period of conventional vertical gel electrophoresis prior to TGGE, the samples migrated into the gel. TGGE was performed in constant voltage (~15 V/cm) mode for 1.5 h. A linear temperature gradient of 30–70°C was usually set using a TGGE apparatus (TG; Taitec Co., Saitama). The gel was removed from the glass plate and stained with silver.

**Computer programs**

The program PCRAna-A1 was newly composed based on a similar algorithm to the previous one (PCRAna), making the program faster and converting the language used from PASCAL to Visual Basic. Briefly, its main function is to calculate Gibbs free energy changes for the primer-binding structures, which are extremely diverse. The process of searching for adequate structures which conform to the constraints imposed have been optimized: geometrically possible, abstract binding structures were first calculated out as ‘skeletons’ and then each partial sequence of the template was checked regarding whether it could match with one of the so-called skeletons as if looking for Cinderella by checking whether one has a foot to fit into a particular shoe (skeleton). This algorithm made the computation of random PCR products faster by more than an order of magnitude in cases of the calculation for a primer of small size (~15 nt). In addition, computation time, in principle, is proportional to the size of the genome sequence, which means it is applicable to even the human genome.

**Poland-H.** The computer program Poland (12) was slightly modified to give Poland-H so that we could easily change parameters required and obtain different modes of output.

Both of these computer programs will soon become publicly available on our web site (http://www.gp.fms.saitama-u.ac.jp/pcranai/).

**Sequencing**

TA cloning (13) was adopted for the preparation of sequencing samples. Sequencing was carried out using a DQS-2000L sequencer (Shimadzu, Kyoto). The DNA fragments of interest were recovered from the DNA band in the gel of random PCR products by performing PCR with specific primers which were designed based on the sequence predicted (see Fig. 3c), which was more efficient and convenient in obtaining significant data than our previous method in which the DNA fragments recovered from the gel were directly cloned into a plasmid and then sequenced. The specifically amplified PCR products were checked by TGGE before sequencing. The sequences obtained
obtained experimentally were checked by comparing them with database sequences of E. coli in DDBJ using BLAST.

RESULTS AND DISCUSSION

Since the predictability of general PCR (in other words random PCR) products had already been established in principle for λ phage genomic DNA (48,502 bp) (2), we examined here whether basically the same program is also effective for genomes which are hundreds of times larger than those of λ phage. We further included a stage where precise one-to-one correspondence between DNA which was experimentally obtained and those which were theoretically predicted could be clearly stated.

The random PCR products obtained with E. coli genomic DNA were subjected to TGGE. Following the procedure depicted in Figure 2, DNA bands of interest which appeared in the gel were further analyzed: (i) their sequences were predicted with the use of a computer program (PCRLana-A1) based on a knowledge of the sequence of the template DNA and primers, together with their size information calculated from their mobility in a gel; (ii) a group of candidates thus generated were further analyzed to select the most probable one with the use of a computer program (Poland-H) which gives a melting map of each sequence. To confirm the prediction results, specific PCR was performed using the proper primers based on a knowledge of the sequence predicted.

Trials 1–3 in Table 1 were performed on three arbitrarily chosen individual DNA bands which appear in the same genome profile (Fig. 3a). Trial 4 was carried out with a DNA band in another genome profile (obtained from a different combination of primer and template) for the sake of generality. One typical result (Trial 3) is shown concretely in Figure 3. In this case, the DNA fragment of ~370 bp, estimated from its mobility, was picked up for sequence prediction. The prediction program (PCRLana-A1) generated a list of candidates as shown in Figure 3. The sequences thus obtained were used to work out melt maps from which we could figure out the initial melting temperature ($T_i$) (see legend to Fig. 3). Combining the theoretical results (size, $\Delta G$, $T_i$ and melting pattern) with the experimental ones, we selected the most probable candidate for the random PCR product of interest as shown in Table 1. In detail, the
The criteria for determining the most probable candidate were as follows: (i) the size difference between the experimental result ($s_e$) and the theoretical one ($s_t$) must be within 10% of $s_e$; (ii) the temperature difference of $T_i$ between the experimental result and the theoretical one must be within ±5°C; (iii) the melting pattern interpreted from the theoretically obtained melt map must be checked to see whether it is similar to those experimentally obtained; (iv) to such a group of candidates which have no significant difference judged from the criteria mentioned above, we further adopted the free energy criterion. We propose that the basic theory of random PCR is that the amount of DNA (2): their dynamics (here that of E.coli genomic DNA) during the course of annealing from the single-stranded state is

\[ \Delta G = \Delta H - T \Delta S \]

will appear in the post-genome sequencing era since it could be used to predict random PCR products obtained from large genomes such as that of E.coli. As a computer program, it can be applied even to the whole human genome sequence (3 Gb), although whether it will be successful or not requires another demonstration. In addition, it is reassuring that the prediction was highly reliable even without considering secondary structure effects (as has already been described in a previous paper (2), considering the local secondary structure in the course of the prediction process is helpful to improve the quality of prediction). This fact of being dispensable is convenient in order to make the computation take less time.

Another important fact found here is that the database genome sequence, with which a wild type of organism cannot be the same, as shown in Table 1, could be effective for the present purpose. The robustness of this prediction method must be related to the nature of the random PCR phenomenon, i.e. primer binding in random PCR is not influenced by mutations occurring inside two primer-binding sites and is relatively less affected by mutations at the binding sites than in the case of conventional PCR due to the rather relaxed primer-binding structures in random PCR (see Fig. 1). This nature is very convenient for such a precious genome sequence to be widely and generally used. A lot of applications of this technology will appear in the post-genome sequencing era since it provides the information of the exact sites on the genome from where random (or general) PCR products are copied, together with the approximate sequences of them without sequencing. One such application is to reinforce the utility of the Genome Profiling database of various species, which is being planned for construction and intended for species identification, by adding the sequence information of DNA fragments (it is evident that only gel electrophoretic band patterns do not give us much of the information on the genome of the relevant organism).

Finally, we must underscore that this experimental system again supported the previous conclusion for solution structures of DNA (2): their dynamics (here that of E.coli genomic DNA) during the course of annealing from the single-stranded state is

\[ r^2 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \]

improving the accuracy of size information, which is obtained through gel electrophoresis, seems to be one of the most convenient ways for raising the prediction score. However, we have to remember that deletion/insertion within the two primer-binding sites will result in an alteration in size without influencing the amount of the fragment. Considering this point of view, the different types of information, $T_i$ and melting pattern, which played a pivotal role in the cases of Trials 2–4, are indispensable for this methodology.

Consequently, the perfect result obtained here must contain some stochastic contribution and should not be made too much of. It may be dangerous to assert, with such a limited number of trials, that our prediction system always offers a correct answer. However, it is evident that the algorithm thus established could be used to predict random PCR products obtained from large genomes such as that of E.coli. As a computer program, it can be applied even to the whole human genome sequence (3 Gb), although whether it will be successful or not requires another demonstration. In addition, it is reassuring that the prediction was highly reliable even without considering secondary structure effects [as has already been described in a previous paper (2), considering the local secondary structure in the course of the prediction process is helpful to improve the quality of prediction]. This fact of being dispensable is convenient in order to make the computation take less time.

Another important fact found here is that the database genome sequence, with which a wild type of organism cannot be the same, as shown in Table 1, could be effective for the present purpose. The robustness of this prediction method must be related to the nature of the random PCR phenomenon, i.e. primer binding in random PCR is not influenced by mutations occurring inside two primer-binding sites and is relatively less affected by mutations at the binding sites than in the case of conventional PCR due to the rather relaxed primer-binding structures in random PCR (see Fig. 1). This nature is very convenient for such a precious genome sequence to be widely and generally used. A lot of applications of this technology will appear in the post-genome sequencing era since it provides the information of the exact sites on the genome from where random (or general) PCR products are copied, together with the approximate sequences of them without sequencing. One such application is to reinforce the utility of the Genome Profiling database of various species, which is being planned for construction and intended for species identification, by adding the sequence information of DNA fragments (it is evident that only gel electrophoretic band patterns do not give us much of the information on the genome of the relevant organism).

Finally, we must underscore that this experimental system again supported the previous conclusion for solution structures of DNA (2): their dynamics (here that of E.coli genomic DNA) during the course of annealing from the single-stranded state is

\[ r^2 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \]

improving the accuracy of size information, which is obtained through gel electrophoresis, seems to be one of the most convenient ways for raising the prediction score. However, we have to remember that deletion/insertion within the two primer-binding sites will result in an alteration in size without influencing the amount of the fragment. Considering this point of view, the different types of information, $T_i$ and melting pattern, which played a pivotal role in the cases of Trials 2–4, are indispensable for this methodology.

Consequently, the perfect result obtained here must contain some stochastic contribution and should not be made too much of. It may be dangerous to assert, with such a limited number of trials, that our prediction system always offers a correct answer. However, it is evident that the algorithm thus established could be used to predict random PCR products obtained from large genomes such as that of E.coli. As a computer program, it can be applied even to the whole human genome sequence (3 Gb), although whether it will be successful or not requires another demonstration. In addition, it is reassuring that the prediction was highly reliable even without considering secondary structure effects [as has already been described in a previous paper (2), considering the local secondary structure in the course of the prediction process is helpful to improve the quality of prediction]. This fact of being dispensable is convenient in order to make the computation take less time.

Another important fact found here is that the database genome sequence, with which a wild type of organism cannot be the same, as shown in Table 1, could be effective for the present purpose. The robustness of this prediction method must be related to the nature of the random PCR phenomenon, i.e. primer binding in random PCR is not influenced by mutations occurring inside two primer-binding sites and is relatively less affected by mutations at the binding sites than in the case of conventional PCR due to the rather relaxed primer-binding structures in random PCR (see Fig. 1). This nature is very convenient for such a precious genome sequence to be widely and generally used. A lot of applications of this technology will appear in the post-genome sequencing era since it provides the information of the exact sites on the genome from where random (or general) PCR products are copied, together with the approximate sequences of them without sequencing. One such application is to reinforce the utility of the Genome Profiling database of various species, which is being planned for construction and intended for species identification, by adding the sequence information of DNA fragments (it is evident that only gel electrophoretic band patterns do not give us much of the information on the genome of the relevant organism).

Finally, we must underscore that this experimental system again supported the previous conclusion for solution structures of DNA (2): their dynamics (here that of E.coli genomic DNA) during the course of annealing from the single-stranded state is

\[ r^2 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \]
Figure 3. One typical result and its logic. From the genome profile, the size and $T_i$ of a DNA band were estimated to be 370 bp and 72°C, respectively. (a) Reduced temperature, in which the effect of a denaturant is converted to temperature (17), is used to show the temperature gradient. (b) Using the appropriate parameters (including the size information), computer program PCRAna-A1 was executed, resulting in a list of candidate sequences. (c) Each candidate sequence was further processed with the program Poland-H to obtain the value of $T_i$ and the melting pattern from the resultant melt map. (d) $T_i$ is shown as the temperature at which >10% local melting occurs. In the melt map, the horizontal axis (left to right) represents the location of the nucleotide along the sequence, the vertical axis probability of each nucleotide not being base paired and the axis towards the back temperature. Each line is drawn to correspond to a particular temperature. The translation of the melting profile is shown beside the melt map. I, double-stranded; C, coiled (or single-stranded); V, terminal melted.
governed mainly by the Watson–Crick base pairing although numerous semi-stable structures of non-Watson–Crick type are already known and potentially contributory. This fact is neither self-evident from the established knowledge nor demonstrable by conventional methodologies, including NMR, due to ignorance of the whole ensemble of structures and their stabilities. Obviously, this is a kind of folding problem, which has been vigorously studied on proteins but far less studied on DNAs and RNAs because of the absence of prominent physical meanings. However, the dynamical solution structures of DNA/RNA have been more and more significant in relation to the function of those molecules (for example the recent finding of the dynamic nature of the lead-dependent ribozyme; 14), especially after the finding of ribozymes from both natural and artificial sources (15,16). Theoretical approaches to this field, therefore, are highly expected and the present success in simulating random PCR products from a long genome sequence should be taken as proof of the rightness of the explanation on the fundamental solution structure of DNA/RNA.

ACKNOWLEDGEMENTS
The authors are grateful to M. Biyani for his technical assistance. This study was supported by a Grant-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan.

REFERENCES