BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 24 2005, pages 4363–4370
doi:10.1093/bioinformatics/bti716

Gene expression

A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3′-end subsequences

Fumihito Miura[1,2], Chihiro Uematsu[3], Yoshiyuki Sakaki[4] and Takashi Ito[1,2,*]

[1]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Japan, [2]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Tokyo, Japan, [3]Central Research Laboratory, Hitachi Ltd, Tokyo, Japan and [4]RIKEN Genomic Sciences Center, Yokohama, Japan

Received on July 27, 2005; revised on October 6, 2005; accepted on October 12, 2005
Advance Access publication October 18, 2005

ABSTRACT

Motivation: In contrast with conventional PCR using a pair of specific primers, some applications utilize a single unique primer in combination with a common primer, thereby relying solely on the former for specificity. These applications include rapid amplification of cDNA ends (RACE), adaptor-tagged competitive PCR (ATAC-PCR), PCR-mediated genome walking and so forth. Since the primers designed by conventional methods often fail to work in these applications, an improved strategy is required, particularly, for a large-scale analysis.

Results: Based on the structure of 'off-target' products in the ATAC-PCR, we reasoned that the practical determinant of the specificity of primers may not be the uniqueness of entire sequence but that of the shortest 3′-end subsequence that exceeds a threshold of duplex stability. We termed such a subsequence as a 'specificity-determining subsequence' (SDSS) and developed a simple algorithm to predict the performance of the primer: the algorithm identifies the SDSS of each primer and examines its uniqueness in the target genome. The primers designed using this algorithm worked much better than those designed using a conventional method in both ATAC-PCR and 5′-RACE experiments. Thus, the algorithm will be generally useful for improving various PCR-based applications.

Availability: The source code of the program is available upon request from the authors or can be obtained from http://itolab.cb.k.u-tokyo.ac.jp/GATC/

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: ito@k.u-tokyo.ac.jp

1 INTRODUCTION

Conventional PCR uses a pair of unique primers to selectively amplify the target sequence. The specificity of amplification is thus achieved by a synergistic effect between the specificities of the two primers. In contrast, some PCR applications use only a single unique primer in combination with a common primer. These include rapid amplification of cDNA ends (RACE) (Frohman *et al*., 1988), PCR-mediated genome walking (Shyamala and Ames, 1989; Riley *et al*., 1990), adaptor-tagged competitive PCR (ATAC-PCR) (Kato, 1997) and so forth. The specificity of the PCR in these applications depends solely on that of the gene-specific primer. Accordingly, they are often plagued with non-specific amplification products derived from so-called 'off-targets.'

These problems may be solved by optimizing the PCR conditions, including the annealing temperature and the concentrations of Mg^{2+}, dNTPs, primers and the template. While methods have been developed for the efficient optimization of PCR conditions (Cobb and Clarkson, 1994), it is impractical to perform such adjustments on each of the thousands of primers used in a large-scale experiment. The most reliable solution that would require no preliminary experiments is to employ a two-step amplification that uses an additional nested gene-specific primer (Gibbons *et al*., 1991). However, this would inevitably double the cost and time. Thus, a reliable method is required to design more specific primers, particularly, for use in a large-scale analysis based on these PCR applications. In addition, the primers should work efficiently under a broad range of experimental conditions so that a single unique condition can be used throughout the large-scale experiments. Such primers would make it possible to optimize the PCR conditions using only a small number of representative primers.

We are developing a genome-wide ATAC-PCR system for the budding yeast *Saccharomyces cerevisiae*. For these experiments, we had designed each gene-specific primer such that the 12mer including its 3′-end (3′-end 12mer) is unique in the yeast genome. Even with these primers, we often encountered off-target products. To determine the etiology of these off-target products, they were cloned and sequenced. As was expected, we found that they were generated by the misannealing of gene-specific primers and, more importantly, that the length of the homologous sequence shared by the primer and the off-target differed from one case to another.

Based on these observations, we assumed that, at least in some instances, the practical specificity may be better determined by the uniqueness of the minimal 3′-end subsequence that exceeds a threshold of duplex stability rather than the uniqueness of entire primer or that of a 3′-end subsequence with a fixed length. We termed such a subsequence as a 'specificity-determining subsequence' (SDSS). Increasing the uniqueness of the SDSS is generally equivalent to its elongation. The longer the SDSS, the more likely it is that the primer functions as a unique one. The shorter the SDSS, the more likely it is that the primer would find off-targets. Primers with a longer SDSS tend to be more AT rich in their 3′-end portions, and the duplex formed by this region is rather unstable,

*To whom correspondence should be addressed.

leading to accurate but inefficient priming. Therefore, a balance should be found between the yield and specificity.

To quantitatively evaluate these issues, we have introduced an intuitive parameter termed as 'association rate' for duplex stability. We have developed an algorithm that calculates the association rate for every 3′-end subsequence of a given primer, identifies its SDSS as the shortest 3′-end subsequence that exceeds a given threshold and examines the uniqueness of the identified SDSS in the target genome, thereby predicting the performance of the primer. We have proved the principle of this strategy in both ATAC-PCR and 5′-RACE experiments using the primers designed based on this algorithm.

2 METHODS

2.1 Preparation of PCR templates

Yeast genomic DNA was prepared from the S288C strain using a Genomic Tip 500 column (QIAGEN) according to the manufacturer's instructions. One microgram of the DNA was digested in 150 µl of 1× K buffer (TAKARA) with 10 units of *Mbo*I (TAKARA) at 37°C for 1 h. Following the heat inactivation of the enzyme at 70°C for 10 min, the reaction was supplemented with 50 µl of ligation buffer [100 mM Tris–HCl (pH 6.9), 10 mM $MgCl_2$, 4 mM ATP and 4 mM DTT] and 100 pmol of adaptor, which was prepared by annealing equimolar amounts of two oligodeoxyribonucleotides: 5′-(PO_3)GAT CCG ATG TGA GCG CCA-3′ and 5′-TGC ACA ATA CTC ACA CAG GAA ACA GCT ATG ACT GCG CTC ACA TCG-3′ (the sequence of the adaptor-specific primer is underlined). Adaptor ligation reaction was started by adding 10 Weiss units of T4 DNA ligase (TAKARA) to the mixture and it proceeded overnight (i.e. >10 h) at 16°C. The reaction was stopped by adding 50 µl of 100 mM Na_2EDTA (pH 8.0) to the mixture. The DNAs were precipitated by isopropanol, dissolved in 100 µl of TE buffer [10 mM Tris–HCl (pH 8.0), 1 mM Na_2EDTA] and further purified using MinElute PCR Purification Kit (QIAGEN) according to the manufacturer's recommendations.

The template for RACE was prepared based on a ligation-anchored PCR (Edwards *et al*., 1991; Troutt *et al*., 1992). Total RNAs were extracted from the S288C cells using the hot phenol method described by Iyer and Struhl (1996). Five microgram of total RNAs was mixed with 0.5 µg of dT_{25} primer in 10 µl of ddH_2O. Following heat denaturation at 70°C for 5 min, the tube was chilled on ice and supplemented with 10 µl of 2× RT solution, comprising 4 µl of 5× RT buffer, 2 µl of 0.1 M DTT, 1 µl of 10 mM of each dNTP, 1 µl of SuperScript II reverse transcriptase and 2 µl of ddH_2O. The tube was sequentially incubated at 25°C for 15 min, 42°C for 60 min and 70°C for 15 min. To degrade the RNAs, 80 µl of 0.01 N NaOH was added to the solution and incubated at 37°C for 30 min. Following neutralization with 100 µl of 100 mM Tris–HCl (pH 8.0), the first-strand cDNA was precipitated with ethanol and dissolved in 10 µl of ddH_2O. To this solution, 5 µl of 10× T4 RNA ligase buffer (TAKARA) and 20 pmol of adaptor 5′-(PO_3)CAT CCA TGG ATC CTC AGC TAG TTA ACT GAG ATA TCG AAT TCC TAT AGT GTC ACC TAA ATC(NH_2)-3′ (the sequence complementary to SP6 primer is underlined) were added and the final volume was adjusted to 25 µl with ddH_2O. The solution was supplemented with 50 units of T4 RNA ligase (TAKARA) and 25 µl of 50% PEG-8000 and incubated at 37°C for >16 h followed by heat inactivation at 70°C for 15 min. Adaptor-anchored cDNAs were ethanol precipitated and dissolved in 50 µl of 1× ExTaq buffer (TAKARA) containing 250 µM of each dNTP and 20 pmol of the SP6 promoter primer (5′-GAT TTA GGT GAC ACT ATA G-3′). Following incubation at 70°C for 10 min, second-strand synthesis was started by adding 5 units of ExTaq DNA polymerase (TAKARA) followed by a 30 min incubation at 70°C. The RACE template DNA was purified with the MinElute PCR Purification Kit with an elution volume of 200 µl. For each PCR, 1 µl of the template solution was used.

2.2 PCR amplification

Each PCR was performed in a 10 µl reaction volume that was composed of 1× PCR buffer (Invitrogen) with 4 mM of $MgCl_2$, 20 µM of each dNTP, 0.2 µM of fluorescence-labeled (at the 5′ terminal) adaptor-specific primer (M13-RV primer 5′-CAG GAA ACA GCT ATG AC-3′ for adaptor-tagged genomic DNA templates, and the SP6 promoter primer for RACE templates), gene-specific primers (Tables 1 and 3), 1 unit of Platinum Taq DNA polymerase (Invitrogen) or TaqHS (TAKARA) and the template DNA prepared as described above. The thermal cycling parameters were as follows: preheating at 94°C for 5 min, 40 (for ATAC-PCR) or 45 (for RACE) cycles of a three-step incubation at 95°C for 20 s, 60°C for 30 s and 72°C for 30 s, followed by a 5 min incubation at 72°C. Amplified PCR products were analyzed by conventional agarose or polyacrylamide gel electrophoresis or by an ABI 3730 Genetic Analyzer (Applied Biosystems) according to the manufacturer's instructions.

2.3 Cloning and sequencing of RACE products

For the analysis of primer specificity in the 5′-RACE, the PCR products were cloned and sequenced: i.e. 5 µl of the PCR products was purified with AMpure reagent (Agencourt) and cloned into the pCR4-TOPO vector (Invitrogen) according to the manufacturer's instructions. For each PCR, 12 positive colonies were picked and their inserts were amplified by PCR with a pair of vector primers, purified with AMpure reagent and sequenced using BigDye version 1.1 cycle sequencing kit (Applied Biosystems).

2.4 Program

To calculate the uniqueness of the SDSS of each primer sequence, we prepared a simple program. This program has two running modes: it evaluates the specificity of a given primer in one mode, whereas in the other mode it extracts candidates for highly specific primers from a given sequence using the SDSS concept. The stability of hybridization was calculated as described in the Results section, using the association rate calculated with the parameters described by SantaLucia (1998), but not ΔG. The parameters for one base-mismatched annealing were taken from Allawi and SantaLucia (1997, 1998a,b,c) and Peyret *et al*. (1999). The program is written in C++ language using the standard template library and the Boost C++ library. The source code for this program is available upon request or can be downloaded from http://itolab.cb.k.u-tokyo.ac.jp/GATC/

3 RESULTS

3.1 Rationale for designing PCR primers by SDSS algorithm

Let $C\mathrm{t}_0$ and $C\mathrm{p}_0$ be the initial concentrations of the template and primer, respectively. Let f be the fraction of the template associated with the gene-specific primer. Then, the concentrations of the template (Ct), primer (Cp), and primer–template complex (Cc) at equilibrium state are defined as follows:

$$C\mathrm{t} = (1-f)\times C\mathrm{t}_0$$
$$C\mathrm{p} = C\mathrm{p}_0 - f\times C\mathrm{t}_0$$
$$C\mathrm{c} = f\times C\mathrm{t}_0.$$

Since the association constant K_{as} is expressed either in terms of ΔG or Ct, Cp and Cc, the following equation is obtained:

$$K_{\mathrm{as}} = \mathrm{e}^{\frac{\Delta G}{RT}} = \frac{C\mathrm{c}}{C\mathrm{p}\times C\mathrm{t}} = \frac{f\times C\mathrm{t}_0}{(C\mathrm{p}_0 - f\times C\mathrm{t}_0)(1-f)\times C\mathrm{t}_0}.$$

Note that ΔG can be calculated as described by SantaLucia (1998) and in associated reports (Allawi and SantaLucia, 1997, 1998a,b,c, Peyret *et al*., 1999). R and T are the gas constant (1.987 cal/K · mol)

Table 1. Oligonucleotide primers used in Figure 2

Name	Sequence[a]	T_m[b]	Frequency of SDSS in genomic sequence	Judgment[c]
Pdr3-1	GACGCATGCTCCTGATACTTCCAATAAT	76.0	1	Specific
Pdr3-2	GACGCATGCTCCTGATACTTCCAATAA	75.8	1	Specific
Pdr3-3	GACGCATGCTCCTGATACTTCCAATA	75.6	1	Specific
Pdr3-4	GACGCATGCTCCTGATACTTCCAAT	76.0	1	Specific
Pdr3-5	GACGCATGCTCCTGATACTTCCAA	75.8	1	Specific
Pdr3-6	GACGCATGCTCCTGATACTTCCA	75.5	2	Not-specific
Pdr3-7	CGACGCATGCTCCTGATACTTCC	76.5	2	Not-specific
Pdr3-8	GCGACGCATGCTCCTGATACTTC	77.0	1	Specific
Pdr3-9	GCGACGCATGCTCCTGATACTT	76.1	1	Specific
Pdr3-10	GCGACGCATGCTCCTGATACT	75.8	1	Specific
Pdr3-11	TGCGACGCATGCTCCTGATAC	76.5	2	Not-specific
Pdr3-12	GTGCGACGCATGCTCCTGATA	76.0	3	Not-specific
Pdr3-13	GTGCGACGCATGCTCCTGAT	76.5	1	Specific
Pdr3-14	GTGCGACGCATGCTCCTGA	76.4	2	Not-specific
Pdr3-15	GTGCGACGCATGCTCCTG	75.4	4	Not-specific
Pdr3-16	GAGTGCGACGCATGCTCCT	76.2	14	Not-specific
Pdr3-17	CGAGTGCGACGCATGCTCC	77.6	7	Not-specific
Pdr3-18	CGAGTGCGACGCATGCTC	75.2	9	Not-specific
Pdr3-19	CCGAGTGCGACGCATGCT	76.6	8	Not-specific
Pdr3-20	TCCGAGTGCGACGCATGC	77.1	5	Not-specific
Pdr3-r	GGAGAACCTCGTCATGTGTATT	70.9	2	Not-specific

[a]Each SDSS determined as described in the text is underlined.
[b]Each T_m was calculated at 1 M NaCl and 0.2 μM primer using the equations described in SantaLucia (1998).
[c]A primer is judged as specific if its SDSS occurs only once in the genome and is judged as not-specific if it does more than twice.

and absolute temperature (333.15 K in our conditions), respectively. Cp_0, in our case, is 2.0×10^{-7} M. Although Ct_0 is usually unknown, it is obviously negligible compared with Cp_0 and can be omitted from the denominator for simplification. Thus, f can be defined as follows:

$$f = \frac{Cp_0 \times K_{as}}{1 + Cp_0 \times K_{as}}.$$

By its definition, the bona fide target is fully matched to the gene-specific primer. Hence, the duplex stability is high and the f value is close to 1. Alternatively, an off-target is matched only to the 3′-end subsequences of the gene-specific primer, and the f value lies in the 0–1 range. The higher the f value, more likely it is that the sequence serves as an off-target. Once the gene-specific primer anneals and primes using a partial homology in the earlier cycles of the PCR, the product serves as an ideal template in the following cycles, leading to prominent off-target products. We had observed this in the off-target products of the ATAC-PCR experiments.

We thus assume that what is practically important is the uniqueness of the shortest subsequence including the 3′-end of the primer that can prime with a substantial efficiency to exceed a threshold in f, but not that of the entire primer sequence. Using the above equation to calculate f for every 3′-end subsequence of a given primer, we can identify the shortest subsequence that exceeds a predetermined threshold value in f. We termed such a sequence as an SDSS.

For instance, the two primers Pdr3-1 and Pdr3-20 shown in Figure 1 differ in the length of SDSS—16mer for the former and 10mer for the latter, when the threshold is set at 0.01. The SDSS of the former primer occurs only once in the budding yeast genome and thus is unique. By contrast, the SDSS of the latter primer occurs five times in the genome. Therefore, the former and the latter are expected to behave as an excellent primer and a poor primer, respectively; we observed that this was the case (Fig. 2). Intriguingly, when judging the two primers based on the uniqueness of their 3′-end 12mers, the former appears worse than

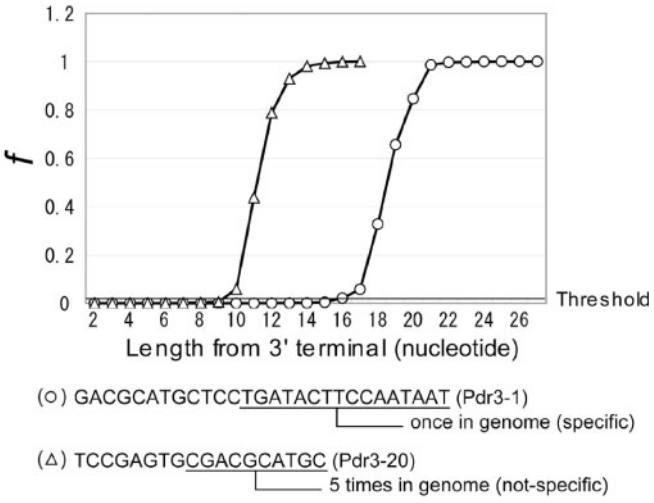

Fig. 1. Calculation of SDSS and its uniqueness in a genome sequence. Two primer sequences are shown as examples. For each 3′-end subsequence of the primer, the f-value was calculated and plotted against its length. The shortest 3′-end subsequence bearing an f-value that exceeds a predetermined threshold is selected as the SDSS of the primer. The frequency of the SDSS in the target genomic sequence indicates the specificity of the candidate primer.

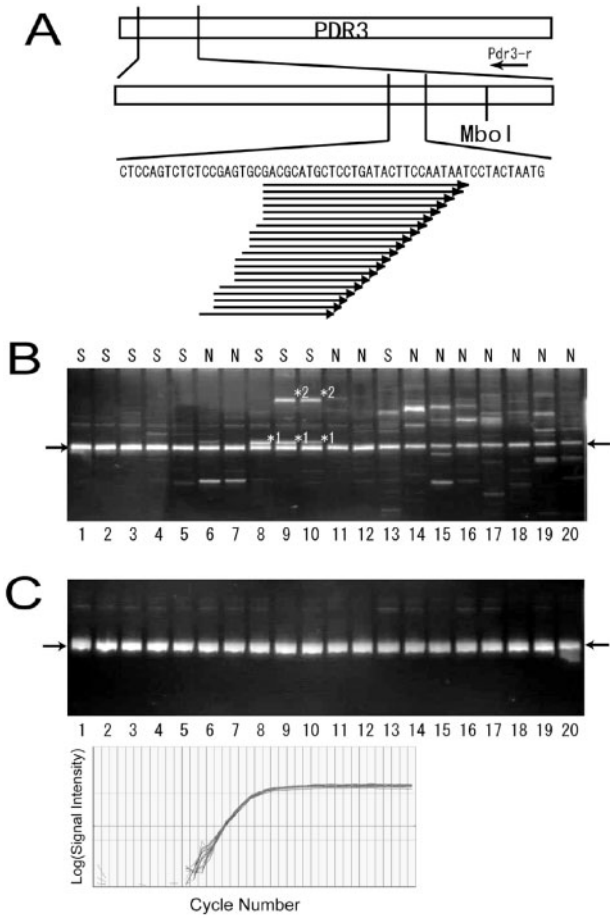

Fig. 2. Correlation between the uniqueness of the SDSS and primer specificity. (**A**) A set of primers (Pdr3-1–Pdr3-20, see Table 1 for details) were designed to tile a segment of the yeast *PDR3* gene for evaluation of the relationship between the SDSS and specificity in PCR. Note that the length of each primer is adjusted so that all of them have approximately the same T_m. An antisense primer (Pdr3-r) was also designed downstream of an *Mbo*I site. (**B**) PCR amplification from an adaptor-tagged genomic DNA template with each of the primers Pdr3-1–Pdr3-20 in combination with a primer specific to the adaptor ligated to the *Mbo*I sites. Each lane is numbered according to the primer ID in Table 1. Arrows indicate the approximate position of expected target bands. 'S' and 'N' indicate that the primer was judged 'Specific' and 'Not-specific,' respectively, by the SDSS algorithm. Asterisks indicate unexpected byproducts, whose identities were determined by sequencing as those including complex mispriming events (Supplementary Fig. 1). (**C**) Control amplification from the genomic DNA template with the same primers as those used in B in combination with the reverse primer Pdr3-r. Electrophoretic patterns of PCR products are also shown (upper panel). Real-time monitoring of PCR amplification with ABI 7000 sequence detection system (Applied Biosystems) indicates that all the primers share comparable amplification efficiency (lower panel). This is presumably because all of the amplicons have a similar length and because all of the 21 primers including Pdr3-r share similar f-values close to 1 under the condition used. The intensity of each amplicon did increase by a factor of ~2 at each cycle of the logarithmic phase of the reaction.

the latter, because the 12mer of the former occurs eight times, whereas that of the latter occurs only once in the genome.

One should note that the basic parameters used here to calculate ΔG by the nearest-neighbor model are determined in the presence of 1 M NaCl, which is much higher than the concentration in the PCR solution. Furthermore, the PCR usually involves the use of KCl rather than NaCl and also $MgCl_2$. While K^+ has been shown to have almost the same effect as Na^+, Mg^{2+} is assumed to have an ~140-fold larger effect than Na^+ (Nakano *et al*., 1999). We set our PCR condition to include 50 mM KCl and 4 mM $MgCl_2$, which can be approximated to be equivalent to a 610 mM concentration of the Na^+ ion. Since the stability of the duplex under these conditions is lower than that predicted under an assumption of the presence of 1 M Na^+, the calculation provides an overestimated f value. Accordingly, the primers designed using these values would have sufficient stringency.

We have developed a simple program which identifies the SDSS of each primer by calculating the f for every 3′-end subsequence and counts its frequency in the target genome sequence, which would serve as a predictor for the performance of the primer.

3.2 Experimental validation of the SDSS algorithm

We designed an experiment using 'tiling' primers for a systematic evaluation of the SDSS algorithm as follows. We designed 20 primers such that their 3′-ends tile a 20 bp segment with a single nucleotide resolution on the second *Mbo*I restriction fragment of *PDR3* gene in budding yeast. The length of each primer is adjusted such that all of them have approximately the same T_m (Table 1, Fig. 2A). On the other hand, we digested the yeast genome with *Mbo*I (5′ GATC 3′) and ligated an adaptor to the cohesive ends. Since an *Mbo*I site is located ~250 bp downstream of the primer sites, we can perform PCR using an adaptor-specific primer and one of the gene-specific primers to obtain products of ~280 bp including the length of the tagged adaptor.

The results of the PCR are shown in Figure 2B. Although all the primers allowed us to obtain PCR products from the adaptor-tagged genomic DNA template, the yield and purity of the bona fide target products are substantially different between the primers. Some PCR products were obtained in high yield and purity, but others were obtained in low yield and were plagued with many off-target products. Notably, all the primers shared a comparable efficiency and specificity in amplification when used in combination with a reverse primer specific to *PDR3* (Pdr3-r) (Fig. 2C). It thus appears that the difference in the specificity of each primer led to variable product yields and purities.

We examined the frequency of the SDSS of each primer in the yeast genome (Table 1). An inverse correlation was observed between the frequency of the SDSS and the performance of the primer—the higher the frequency of the SDSS, the poorer the primer performance tended to be. Although the correlation was not complete, we can roughly predict the performance of the primer based on the uniqueness of the SDSS and can, at least, eliminate apparently poor primers.

3.3 Application of the SDSS algorithm to a large-scale PCR primer design for adaptor-tagged templates

We examined the performance of the SDSS algorithm in the design of gene-specific primers for ATAC-PCR of 96 ORFs from the budding yeast genome. We designed a primer set covering these 96 ORFs based solely on the uniqueness of the 12 nt subsequence including the 3′ end (i.e. the 3′-end 12mer), since most 12mers occur uniquely in the yeast genome (Supplementary Table 1). We also designed another primer set using the SDSS algorithm with the

Table 2. Comparison of the primer extraction algorithms[a]

Primer selection policy	3′-end 12mer	SDSS
Specific amplification	8 (8%)	54 (56%)
Non-specific amplification	81 (84%)	42 (43%)
No amplification	7 (7%)	0 (0%)
Total	96	96

[a]Data in Figure 3A and B are summarized. A primer is judged as non-specific if the intensity of the off-targets exceeded 5% of that of the bona fide target.

threshold of f set at 0.01 (Supplementary Table 2). For some ORFs, we could not design any 'unique' primer fulfilling the requirements. In these cases, we allowed primers that may generate off-target products longer than 350 bp excluding the length of the tagged adaptor. All of these primers were designed to have a similar T_m and were tested using the adaptor-tagged genomic DNA as the template under a single defined condition without any *ad hoc* adjustment of the reaction condition.

Target bands were detected in >90% of the cases, regardless of the primer design algorithm. However, in >80% of the cases using the primers based on the 3′-end 12mer specificity algorithm, non-specific or off-target bands were detected; the amplified products contained bands derived from both bona fide targets and off-targets (Table 2). Note that we judged a case as non-specific if the intensity of the off-targets exceeded 5% of that of the bona fide target. In contrast, the use of the SDSS algorithm substantially reduced the occurrence of such cases; off-targets were observed in 43% of the cases (Table 2).

Furthermore, visual inspection of individual electropherograms suggested that the products obtained by the 3′-end 12mer specificity algorithm tend to have more off-targets than those obtained by the SDSS algorithm (Fig. 3A and B). For the quantitative evaluation of these data, we calculated the ratio of the signal intensities between the bona fide target and all other off-targets appearing in a window ranging from 60 to 350 bp. The signal-to-noise ratio (S/N) of each PCR was plotted against the signal intensity of the bona fide product (Fig. 3C). The results clearly demonstrated the superior performance of the SDSS algorithm in designing specific assays; the SDSS primers gave higher bona fide signals and S/Ns than those designed based on the 3′-end 12mer specificity algorithm.

3.4 Application of the SDSS algorithm to RACE primer design

Next, we applied the SDSS algorithm to the design of primers for 5′-RACE, because it is one of the most frequently used techniques in cDNA cloning. Even with the complete genome sequence data, it is currently impossible to predict the transcriptional start sites. Thus, these sites have to be determined by various experimental approaches, among which 5′-RACE represents one of the most popular and reliable ones. In RACE, the specificity of the primer *per se* is not the sole determinant of the 'practical' specificity, because the abundance of the targets differs drastically from one case to another. Consequently, the less abundantly the target is expressed, the more stringently the primer should be designed. Even though the f between a primer and an off-target is 0.01, if the off-target is expressed 100-fold more abundantly than the bona fide target, the primer should not be used for RACE. In other words,

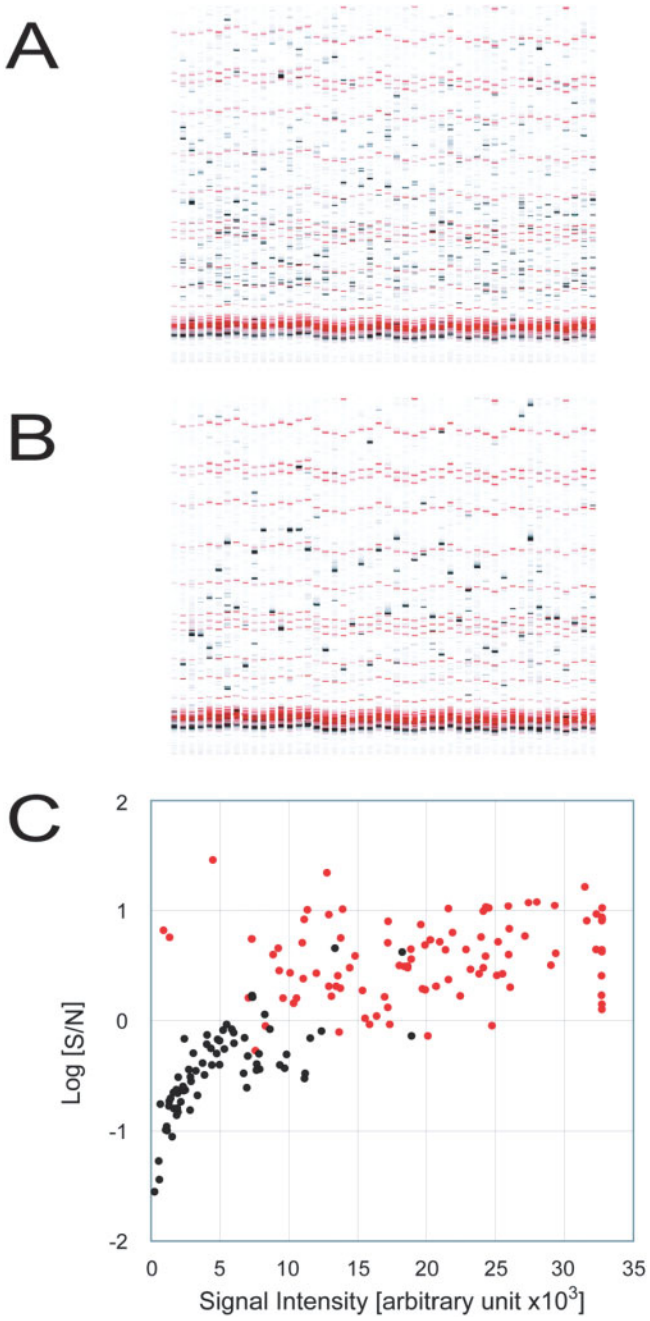

Fig. 3. Improvement of ATAC-PCR using the SDSS algorithm. Two sets of gene-specific primers, one designed based on the uniqueness of the 3′-end 12mer (**A**) and the other by using the SDSS algorithm (**B**), were used for the ATAC-PCR of 96 ORFs from yeast genomic DNA. PCR products and size standards are indicated by black and red channels, respectively. (**C**) The ratios S/N were plotted against the signal intensities of bona fide targets. Black and red circles indicate the results obtained in A and B, respectively.

we have to take the relative expression level between the bona fide target and the off-targets into account. Thus, one should ideally use the following expression instead of just f:

$$f' = f \times (\text{relative expression}).$$

Table 3. Oligonucleotide primers used for the 5′-RACE in Figure 4

Primer name	Target ORF	Primer sequence (SDSS underlined)	Frequency of SDSS in genomic sequence
CapSp_01	YAL001C	TCCTTTTTTGTGTATCCCGTTAATAATGT	1
CapSp_02	YAL007C	CCCCACTCCAAACGATTTTAATAAAAAG	1
CapSp_03	YAL008W	GCATGCTTGTAATACCGACATACATA	1
CapSp_04	YAL013W	CCTGCGTCACTGGATATACAGTA	1
CapSp_05	YAL014C	ACAACTGCTGTTGCTGGTTAATAAATA	1
CapSp_06	YAL018C	CTAAGCACACCTCCACTAAATGATTAT	1
CapSp_07	YAL023C	GCGTTGAACAGTCTCATTTTAACATAAT	1
CapSp_08	YAL024C	CTCCCTAATACACCATCGAAATCTATAG	1
CapSp_09	YAL025C	TGACCATTAGGTGCCTTAATTCTATG	1
CapSp_10	YAL026C	GGTCGTCATGGTTATTGCTCATAAATAG	1
Cap12_01	YAL001C	TGCCCAGTTACCTGCGCC	467
Cap12_02	YAL007C	GGTAACCCACAGCCAGGG	129
Cap12_03	YAL008W	GCCAGCACCGCCTAACCC	18
Cap12_04	YAL013W	GCTCGTCCCTGGCGTCGG	253
Cap12_05	YAL014C	GGCAGCGGCGACTCCTCG	34
Cap12_06	YAL018C	GCCCAATAGGGCCCAGCA	162
Cap12_07	YAL023C	TGGCACACAGAGCGCGGA	263
Cap12_08	YAL024C	CGCGGGCAGGTCTGCACT	41
Cap12_09	YAL025C	GGCAGGAGTGTGCGCTCT	158
Cap12_10	YAL026C	CCGGGAGGCTTGACTGCT	26

The primers named as 'CapSp∼' were designed to tolerate 10-fold abundance of off-targets against the bona fide targets or were designed using the threshold value at 0.001 instead of 0.01 to define the SDSS. The primers named as 'Cap12∼' were designed so that each 3′-end 12mer is unique in budding yeast genomic sequence.

However, it is usually impossible to know the relative expression level between two different genes. Here we used a tentative relative expression value of 10 to design gene-specific primers for the 5′-RACE of 10 ORFs in the yeast chromosome 1. We also designed another set of primers for these 10 ORFs such that their 3′-end 12mer sequences are unique in the budding yeast genome (Table 3).

These two sets of primers were used for the 5′-RACE, and the products were resolved by gel electrophoresis (Fig. 4A). Since it is difficult to judge whether the 5′-RACE has worked solely from the electrophoretic patterns, we cloned the products, randomly picked 12 clones and sequenced them for identification. As summarized in Figure 4B, the primers designed using the SDSS algorithm worked strikingly better than those designed using the 3′-end 12mer specificity algorithm. More than 90% of the products obtained using the SDSS primers were derived from the bona fide targets (Fig. 4B). Even in the cases where no distinct bands were observed on the gel, bona fide targets with different ends were detected. These were presumably due to truncated reverse transcription. In contrast, the purity of the products obtained by using the other primer set was rather poor (∼10%) (Fig. 4B). These results clearly demonstrate the efficiency of the SDSS algorithm in designing primers for RACE, particularly for less abundant transcripts such as the 10 ORFs examined above.

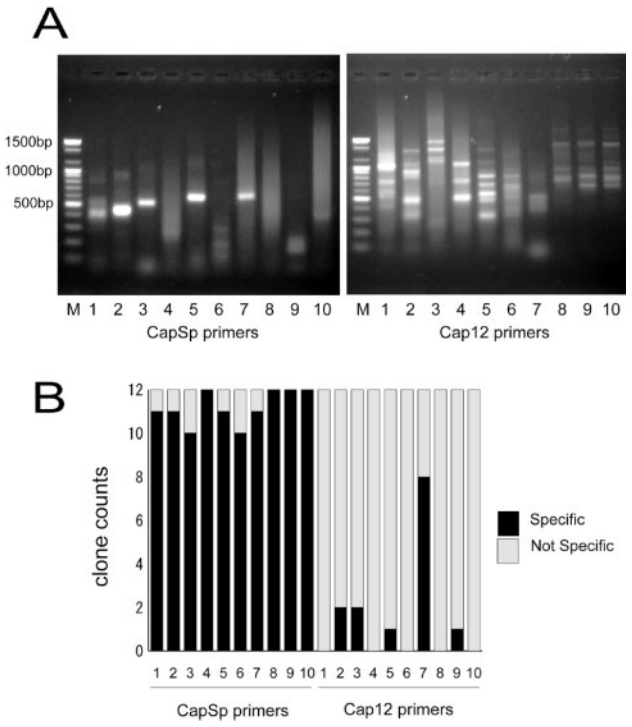

Fig. 4. Improvement of 5′-RACE by the SDSS algorithm. (**A**) The 5′-RACE products obtained using the primers designed based on the uniqueness of the SDSS (left) and the 3′-end 12mer (right) were separated by gel electrophoresis. (**B**) The 5′-RACE products shown in (A) were cloned. For each target, twelve clones were sequenced to examine how effectively the bona fide target is cloned. Closed and gray boxes indicate the true and false targets, respectively.

4 DISCUSSION

A number of programs and web-based services are available for designing PCR primers. Some of them use primer specificity as a parameter to select gene-specific primers from many candidate primers. The specificity of a primer may be defined as 'the ability of a primer to hybridize to no sequences other than a bona fide target,' and some objective parameters were employed for its

description. For instance, 'secondary binding sites' including mismatched hybridization (Haas *et al*., 1998) or higher similarity (Rozen and Skaletsky, 2000) are considered simply using the entire primer sequence. An implicit assumption underlying these may be that stable hybridization of a primer with the template is a prerequisite for priming by DNA polymerase. However, this is not always the case; DNA polymerases are known to prime from an incomplete duplex formed between the primer and the template. Accordingly, some other programs paid attention to the 3′-end portion of the primer. These calculate the frequency of 6mers (Oligo program based on Rychlik and Rhoads, 1989; Rychlik *et al*., 1990; Hyndman and Mitsuhashi, 2003) or the stabilities of the 8 bp (Haas *et al*., 1998), 6 bp (Chen *et al*., 2003), or 5 bp segments (Rozen and Skaletsky, 2000) of the 3′-terminal subsequences. Note that these algorithms use either the entire primer sequence or 3′-subsequence of a fixed length for the calculation.

Although the use of a fixed length makes the calculation much simpler, the length to be considered should be, in principle, different from one primer to another, depending on the properties of the sequences. A primer whose 3′-end portion can form a duplex more stable than −11 kcal/mol with the template was shown to support the priming by DNA polymerases (Rychlik, 1995). The minimum length of 3′-end sequences that matches to the template varies from 8 bp for G-C rich one to 13 bp for A-T rich one, under the conditions used by Rychlik (1995). In addition, our analysis of the off-target products of the ATAC-PCR revealed that the length of homology between the off-target sequence and the 3′-end subsequence of the primer is not fixed but quite variable (data not shown). These observations not only reinforce the importance of 3′-end subsequence but also indicate that the length of the subsequence to be considered should be adjusted depending on the primer sequence. How can we determine the length to be considered?

For this purpose, we introduced the concept of the SDSS or the shortest 3′-end subsequence of a primer that exceeds the threshold in association rate. We use the frequency of the SDSS in the target genome as a predictor of the primer specificity. To the best of our knowledge, this is the first algorithm to rationally define the length of the 3′-end subsequence to be considered. Indeed, the efficiency of the SDSS algorithm was demonstrated in both ATAC-PCR and 5′-RACE, compared with an algorithm based on the uniqueness of a 3′-end subsequence with a fixed length. Hence, the SDSS algorithm will improve the performance of various PCR applications, particularly, those using a single unique primer with a common primer.

While the SDSS at $f = 0.01$ can be as short as 7 nt under the conditions that we used, the DNA polymerase requires at least 8 bp duplex to start extending from the primer (Rychlik, 1995). Accordingly, if the SDSS of a primer is 8 nt or longer, every sequence identical to the SDSS should be regarded as a potential off-target site for the primer. On the other hand, if the SDSS of a primer is 7 nt, this algorithm overestimates the frequency of its off-target sites, because approximately three-quarters of the genomic sequences identical to the 7mer SDSS would form only 7 bp duplexes with the primer and hence fail to support the priming by the enzyme. One may thus relax the criteria by considering only the sites that can form duplexes longer than 8 bp with the primer. Such adjustment would not be necessary for the yeast or larger genomes, because even the 3′-end 8mer, which is one-base extended to the 5′-end from the 7 nt SDSS, likely occurs at such a high frequency that the primer cannot be accepted. However, the adjustment may be useful in some instance when applying the algorithm to much smaller genomes such as those of viruses.

Based on its nature, our approach is most suitable to organisms with fully determined genome sequences. However, this approach would be useful even in organisms with partial genome sequence data, because it can at least exclude poor primers.

Even with the primers designed using the SDSS algorithm, we occasionally encounter prominent off-target products in ATAC-PCR (Fig. 2B, asterisks); the generation of these products was found to involve complex patterns of GT and GA mismatches and gapped annealing in the SDSSs (Supplementary Fig. 1). This is also the case for most of the off-targets of 5′-RACE (Fig. 4B). These complex off-targets may be eliminated by optimizing the PCR conditions using the modified Taguchi method (Cobb and Clarkson, 1994). We successfully applied this method to find a condition to reduce some off-targets in ATAC-PCR, but this led to the enhancement of other off-targets (data not shown). Thus, it appears that, in addition to experimental optimization efforts, the integration of a more sophisticated algorithm for off-target search with the SDSS algorithm is necessary to prevent the formation of complex off-target products. Such improvements may provide a more robust SDSS strategy to design highly specific primers that can be used for genomes with a much higher complexity.

ACKNOWLEDGEMENTS

This work was supported by the Bioinformatics Research and Development (BIRD) project of the Japan Science and Technology Agency (JST), Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Industrial Science and Technology Program of the New Energy and Industrial Technology Development Organization (NEDO), Japan.

Conflict of Interest: none declared.

REFERENCES

Allawi,H.T. and SantaLucia,J.,Jr (1997) Thermodynamics and NMR of internal G/T mismatches in DNA. *Biochemistry*, **36**, 10581–10594.

Allawi,H.T. and SantaLucia,J.,Jr (1998a) Nearest neighbor thermodynamic parameters for internal G/A mismatches in DNA. *Biochemistry*, **37**, 2170–2179.

Allawi,H.T. and SantaLucia,J.,Jr (1998b) Nearest-neighbor thermodynamics of internal A/C mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, **37**, 9435–9444.

Allawi,H.T. and SantaLucia,J.,Jr (1998c) Thermodynamics of internal C/T mismatches in DNA. *Nucleic Acids Res*., **26**, 2694–2701.

Chen,S.H. *et al*. (2003) Primer design assistant (PDA): a web-based primer design tool. *Nucleic Acids Res*., **31**, 3751–3754.

Cobb,B.D. and Clarkson,J.M. (1994) A simple procedure for optimizing the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Res*., **22**, 3801–3805.

Edwards,J.B. *et al*. (1991) Oligodeoxyribonucleotide ligation to single-stranded cDNAs: a new tool for cloning 5′ ends of mRNAs and for constructing cDNA libraries by *in vitro* amplification. *Nucleic Acids Res*., **11**, 5227–5232.

Frohman,M.A. *et al*. (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA*, **85**, 8998–9002.

Gibbons,I.R. *et al*. (1991) A PCR procedure to determine the sequence of large polypeptides by rapid walking through a cDNA library. *Proc Natl Acad. Sci. USA*, **88**, 8563–8567.

Haas,S. *et al*. (1998) Primer design for large scale sequencing. *Nucleic Acids Res*., **26**, 3006–3012.

Hyndman,D.L. and Mitsuhashi,M. (2003) PCR primer design. *Methods Mol. Biol*., **226**, 81–88.

Iyer,V. and Struhl,K. (1996) Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **93**, 5208–5212.

Kato,K. (1997) Adaptor-tagged competitive PCR: a novel method for measuring relative gene expression. *Nucleic Acids Res*., **25**, 4694–4696.

Nakano,S. *et al*. (1999) Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Res*., **27**, 2957–2965.

Peyret,N. *et al.* (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A/A, C/C, G/G, and T/T mismatches. *Biochemistry*, **38**, 3468–3477.

Riley,J. *et al*. (1990) A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res*., **18**, 2887–2890.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.

Rychlik,W. and Rhoads,R.E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and *in vitro* amplification of DNA. *Nucleic Acids Res*., **17**, 8543–8551.

Rychlik,W. *et al.* (1990) Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucleic Acids Res*., **18**, 6409–6412.

Rychlik,W. (1995) Priming efficiency in PCR. *Biotechniques*, **18**, 84–90.

SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.

Shyamala,V. and Ames,G.F. (1989) Genome walking by single-specific-primer polymerase chain reaction: SSP-PCR. *Gene*, **84**, 1–8.

Troutt,A.B. *et al.* (1992) Ligation-anchored PCR: a simple amplification technique with single-sided specificity. *Proc. Natl Acad. Sci. USA*, **89**, 9823–9825.