FastPCR Software for PCR Primer and Probe Design and Repeat Search

Ruslan Kalendar1,2* • David Lee3 • Alan H. Schulman1,4

1 MTT/BI Plant Genomics, Institute of Biotechnology, P. O. Box 65, FIN-00014 Helsinki, University of Helsinki, Finland
2 PrimerDigital Ltd, FIN-00790 Helsinki, Finland
3 John Bingham Laboratory, National Institute of Agricultural Botany, Huntingdon Road, Cambridge CB3 0LE, UK
4 Biotechnology and Food Research, MTT Agrifood Research Finland, FIN-33100 Jokioinen, Finland

Corresponding author: * ruslan.kalendar@helsinki.fi

ABSTRACT

Reproducible and target-specific polymerase chain reaction (PCR) amplification relies on several interrelated factors of which primer design is central. Here, we describe new free bioinformatics software, the FastPCR which was developed, and continues to be updated, based on detailed experimental studies of PCR efficiency for the optimal design of primers and probe sequences and for repeat searching. This software forms an environment of integrated tools, which provides comprehensive facilities for designing primers for most PCR applications including multiplex and self-reporting fluorescent systems. FastPCR consists of a data editor, build commands for probe and primer design, and automation tools. The software selects the best primers with the widest range of melting temperatures, which allows designing qualified primers for all PCR tasks. The “in silico” PCR primer or probe searching includes comprehensive individual primers and primer pair analysis tests. FastPCR utilizes combinations of normal and degenerate primers for all tools. The melting temperature calculation is based on nearest neighbour thermodynamic parameters starting with multiple nucleic acid or protein sequences. It performs efficient and complete detection of various repeat types with visual display. FastPCR is able to perform repeat searches for a single sequence or for comparisons of two sequences. The program includes various bioinformatics tools for analysis of sequences with GC or AT skew, GC content, and purine-pyrimidine skew, and considers linguistic sequence complexity. It can generate random DNA sequence, make restriction analysis, and supports the clustering of sequences and consensus sequence generation, as well as sequence similarity and conservancy analyses.

Keywords: in silico PCR, optimal annealing temperature, secondary (non-specific) binding test

Abbreviations: LTR, long terminal repeat; LUX, light upon extension; MITE, miniature inverted repeat transposable element; PCR, polymerase chain reaction; RT-PCR, real-time PCR; S-SAP, sequence-specific amplification polymorphism; SSR, simple sequence repeat

CONTENTS

INTRODUCTION .......................................................................................................................................................................................... 1
Algorithm and implementation ........................................................................................................................................................................ 2
PCR primer design generalities ..................................................................................................................................................................... 2
Melting temperature (Tm) calculation ...................................................................................................................................................... 3
Linguistic complexity analysis .................................................................................................................................................................. 4
Primer quality determination .................................................................................................................................................................... 4
Hairpin (loop) and dimer formation ...................................................................................................................................................... 5
Calculation of optimal annealing temperature ........................................................................................................................................ 5
PCR amplification protocol ................................................................................................................................................................... 6
The secondary (non-specific) binding test (alternative amplification) ................................................................................................ 6
Simple sequence repeat (SSR) loci search ............................................................................................................................................. 7
In silico PCR .............................................................................................................................................................................................. 7
Group-specific PCR primers (family-specific primer set) and unique PCR ......................................................................................... 8
Real-time PCR primers and probe design ................................................................................................................................................ 8
Primer analysis ....................................................................................................................................................................................... 9
Sequence manipulation tools ................................................................................................................................................................. 9
Alignment algorithm and repeat search .................................................................................................................................................... 9
An efficient tool for discovering LTR retrotransposons ....................................................................................................................... 9
Simple sequence repeat (SSR) loci search ............................................................................................................................................. 11
DISCUSSION .......................................................................................................................................................................................... 12
CONCLUSION ...................................................................................................................................................................................... 12
AVAILABILITY ..................................................................................................................................................................................... 12
REFERENCES ...................................................................................................................................................................................... 13

INTRODUCTION

The polymerase chain reaction (PCR) is fundamental to molecular biology and a key practical molecular technique for the clinical laboratory. However, the efficiency of the method is dependent on identifying unique primer sequences and designing PCR-efficient primers. Primer design is a critical step in all types of PCR methods to ensure specific and efficient amplification of a target sequence (Rychlik et al. 1990). Even today, with many online and commercial bioinformatics tools, primer design for their application in PCR is still not convenient or practical for routine use.
The implementation of PCR technology for new and very specific applications has made it necessary to develop new criteria for PCR primer and probe design, considering alternatives such as RT-PCR, qPCR, group-specific or unique PCR primer design, single primer PCR, combinations of multiple primers in multiplex PCR, amplification of simple nucleotide repeats by surrounding PCR, primer design of PCR primers for direct, inverted or tandem repeats as in RapD (Welsh and McClelland 1990; Williams et al. 1990) and IRAP/REMAP (Kalendar et al., 1999, Kalendar and Schulman 2006), and probe design (dual-labeled oligonucleotide probes - TaqMan® or molecular beacons and microarray long oligonucleotides).

In developing FastPCR, our aim was to develop a practical and easy-to-use professional software for routine manipulation and analysis of sequences, PCR primer and probe design, in silico PCR and oligonucleotide analyses, primer analysis and clustering, routine primer searches, DNA restriction enzyme analysis, sequence alignment, repeat searching and all necessary tools related to these tasks. Some important bioinformatics tools were included for analysis of sequences with varying GC content and nucleotide ratios; these include analyses of purine-pyrimidine, GC, and AT composition, melting temperature, linguistic sequence complexity, consensus sequences, sequence pair similarity, and sequence conservation, as well as a routine for generation of random DNA sequences.

The program is based on our experimental data for efficient PCR amplification, which were translated into algorithms in order to design combinations of primer pairs for optimal PCR amplifications. In addition, we have incorporated calculations for optimal annealing temperature parameters and PCR amplification efficiency for each primer, as well as several other tools for sequence manipulation and data format conversion.

This software, FastPCR, has been successfully used throughout the scientific community in a wide range of PCR and probe applications, as well as in repeat searches and analyses. Since 1999, this software has been applied in various projects, cited in many scientific articles (over 240 hits using Google Scholar search in June 2009: 45 articles at ScienceDirect, 52 articles at Wiley InterScience, 33 articles at SpringerLink, 15 articles at BioMed Central, 5 articles at Oxford Journals, 2 articles at Nature Publishing Group), patents, PhD theses (some of them are shown on our web site: http://www.biocenter.helsinki.fi/bi/Programs/citation.html) and over 1300 downloads of the installation file per month with 3000 unique visitors around world. The program code is regularly updated.

Algorithm and implementation

The FastPCR software is written in Microsoft Visual Studio 6.0 and compiled to an executable file that, after installation, can be used with any version of Microsoft Windows. Although currently FastPCR is not officially supported on the Linux platform, users report being able to run FastPCR successfully on Linux using WINE. For Linux and Mac, WINE serves as a compatibility layer for running Windows programs and is a completely free alternative implementation of the Windows API also for use of native Windows DLLs. For Linux installation, it is not necessary to compile the source code; entering 'wine msiexec /i FastPCR.msi' will install the software without any problem. Users who want to use FastPCR on the Mac platform are encouraged to use the Virtual PC Windows Emulator from Microsoft.

The program takes either a single sequence or accepts multiple separate DNA, RNA, or protein sequences in any of various formats: FASTA, tabulated, Excel sheet, Word table, EMBL, MEGA, GenBank, Msf alignments, Dialign, simple alignment, Blast Queue web alignments results, which are output as ASCII text, XML, or Excel. Users may paste or load files of the target sequence(s) into the sequence input text RTF editor. The segmented output of the primer or probe design includes a list of primers, a set of primer pair sequences with their theoretical PCR products, and, for multiplex PCR, the result of the calculation of multiple PCR primers for given target sequences. In addition, the output shows the optimal annealing temperature for each primer pair and the size of PCR product and complete information for each designed primer or multiplex PCR product set.

The PCR primer design algorithm generates a set of primers with a high likelihood of success using most amplification protocols. All PCR primers designed by FastPCR can be used for PCR or sequencing experiments. The calculation of the primer melting temperature (Tm), dimer detection and other oligonucleotide features (GC content, linguistic sequence complexity, molar extinction coefficient, molecular weight, mmole and mg per OD260), and dilution and resuspension can be also tested in a Java applet on our Web site.

PCR primer design generalities

Primer design is one of the key steps for successful PCR. The major parameters of primers include their nucleotide structure and melting temperature. Usually, PCR primers are 18-35 bases in length (PCR primer or probe length can vary from 15 to 500 bases) and should be designed such that they have complete sequence similarity to the desired target fragment to be amplified. The parameters controllable by the user are: primer length; melting temperature with different nearest neighbour thermodynamic parameters or simple formulae; pyrimidine-purine linguistic complexity; primer GC content; GC 3’ end terminal enforcement; 3’ or 5’ end nucleotide composition for degenerate primers; poly-purine (T, C) or poly-pyrimidine (A, G) stretches; and limits to repeats of identical bases.

The other main parameters used in FastPCR for primer selection are: the general nucleotide structure of the primer such as linguistic complexity (nucleotide arrangement and composition); specificity; the melting temperature of the whole primer and the melting temperature at the 3’ and 5’ termini; a self-complementarity test; secondary (non-specific) binding. Optionally, the software can dynamically optimize the best primer length for entered parameters. All PCR primer (probe) design parameters are flexible and changeable according to the specifics of the analyzed sequence and the task. Primer pairs are analyzed for cross-hybridization, the specificity of both primers, and optionally for similar melting temperature. Designed primers with balanced melting temperatures (within 1-6°C of each other) are desirable but not mandatory. The default primer design selection criteria are shown in Table 1.

In many cases it is necessary to use a predesigned single primer or a list of primers or probes. The program accepts a list of predesigned oligonucleotide sequences for checking the compatibility of each primer with a newly designed primer or probe. The software allows predesigned oligonucleotides that feature as part of the final PCR primer design

### Table 1: Default primer design selection criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Default</th>
<th>Ideal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (nt)</td>
<td>20 – 25</td>
<td>22 – 35</td>
</tr>
<tr>
<td>Tm range (°C)</td>
<td>55 – 65</td>
<td>60 – 68</td>
</tr>
<tr>
<td>Tm* at 3’-end, 12 bases</td>
<td>30 – 48</td>
<td>41 – 45</td>
</tr>
<tr>
<td>Tm* at 5’-end, 12 bases</td>
<td>30 – 45</td>
<td>38 – 45</td>
</tr>
<tr>
<td>GC (%)</td>
<td>45 – 65</td>
<td>50</td>
</tr>
<tr>
<td>3’-end composition (5’-nnn-3’)</td>
<td>sws, wss, ss, wss, ssa, ssw</td>
<td>swa</td>
</tr>
<tr>
<td>5’-end composition (5’-nnn-3’)</td>
<td>nnn, tnn</td>
<td>tns</td>
</tr>
<tr>
<td>Sequence linguistic complexity*</td>
<td>&gt;80%</td>
<td>&gt;90%</td>
</tr>
<tr>
<td>Purine-pyrimidine linguistic complexity*</td>
<td>&gt;80%</td>
<td>&gt;90%</td>
</tr>
<tr>
<td>Degenerate bases at 3’ end, 12 bases</td>
<td>&lt; 3</td>
<td>0</td>
</tr>
<tr>
<td>‘Primer Quality’</td>
<td>80</td>
<td>&gt;90</td>
</tr>
</tbody>
</table>

* Nearest neighbour thermodynamic parameters (Allawi and SantaLucia 1997).

* Sequence linguistic complexity measurement was performed using the alphabet-capacity 1-gram method.
results. The software automatically checks the primer sequence location (with local alignment) on a target sequence and adds compatible primers to a list of selected primers. The user can vary the product size or design primer pairs for the whole sequence without specifying parameters by using default or pre-designed parameters. The pre-designed parameters are specified for different situations: for example, sequences for low GC content, or long distance PCR, or degenerate sequences, or for manual input.

The program is able to generate either long oligomers or PCR primers for the amplification of gene-specific DNA fragments of user-defined length. Up to now, several primer/oligo design programs have been developed (Rubin and Levy 1996; Rozen and Skaletsky 2000; Lexa and Valle 2003; Gadberry et al. 2005; Yang et al. 2006; Fredslund et al. 2007; Bekaert and Teeling 2008). All of them are specialized for either the design of PCR primers or oligomers. Our FastPCR software provides the more flexible approach of designing primers for many applications with quality and speed and will check if either primers or probes have secondary binding sites that may give rise to an additional PCR product. The evaluation of potential secondary binding sites for each primer is performed with local repeat dataset sequences. The selection of the optimal target region for the design of long oligomers is performed in the same way as for PCR primers. The basic parameters in primer design are also used as a measure of the oligomer quality. However, the thermodynamic stability of the 3’ and 5’ terminal bases and central part of the oligomer are evaluated.

Melting temperature (Tm) calculation

The Tm is defined as the temperature at which half of the strands are in the double-helical state and half are in the
“random-coil” state. The Tm for oligonucleotides with normal or degenerate (mixed or “wobble”) nucleotide combinations are calculated in the default setting using nearest neighbour thermodynamic parameters (Sugimoto et al. 1996; Allawi and SantaLucia 1997; Gilson et al. 1997; SantaLucia 1998; Peyret et al. 1999; Bonmarito et al. 2000; Novere 2001). The GC content of an oligonucleotide is the most important factor that influences the Tm value (Fig. 1). This picture shows that the melting temperature, calculated by nearest neighbour thermodynamic parameters, is clearly correlated (>0.96) with the GC content for windows of 20, 30, and 50 oligonucleotides in length.

The melting temperature for mixed bases is calculated by averaging nearest neighbour thermodynamic parameters – enthalpy and entropy values - at each mixed site; extinction coefficients calculated by nearest neighbour thermodynamic parameters, which are calculated by nearest neighbour thermodynamic parameters or simple non-thermodynamic parameters (Sugimoto et al. 1999; Bommarito et al. 2008). In Table 2 and 3 are shown the nearest neighbour thermodynamic parameters, which are the enthalpy and entropy values for pure and mixed nucleotides. The first nucleotide in N2,N2 is shown in the horizontal column and second nucleotide N2,N2 in the vertical column. FastPCR allows the choice of other nearest neighbour thermodynamic parameters or simple non-thermodynamic Tm calculation formulae. For non-thermodynamic Tm calculation for oligonucleotides, we suggest to use the simple formula: Tm=2(A + T) + 4(G + C) (for short <15 bases) or Tm=64.9 + 41(G + C - 16.4)/L (for primers longer than 15 bases). Mismatched pairs can be taken into account since the parameters provide for DNA/DNA duplexes and dangling ends, the unmatched terminal nucleotides (Novere 2001). The melting temperature for primer (probe) self or cross dimers and for in silico PCR experiments with oligonucleotides with mismatches to the target is calculated using values for the thermodynamic parameters for a nucleic acid duplex.

### Linguistic complexity analysis

The sequence analysis complexity calculation method can be used to search for conserved regions between compared sequences, for the detection of low-complexity regions including simple sequence repeats or imperfect direct or inverted repeats. Linguistic complexity measurements are performed using the alphabet-capacity l-gram method (Gabriel and Bolshev 1995; Polony and Potapov 2004) with a sliding window of 10 to 2000 nt, or the whole primer length. The profile is constructed by averaging the complexity values of all sequences within a set in each window. The complexity values are converted to a percentage value, with 100% being the highest level of sequence complexity.

### Primer quality determination

Our experimental data showed that the nucleotide composition and the melting temperature of 12 bases at the 3’ of the primers are important factors for PCR efficiency. The melting temperature of the 12 base terminus is preferably calculated by nearest neighbour thermodynamic parameters (Allawi and SantaLucia 1997; SantaLucia 1998). The composition of the sequence at the 5’-terminus is less important, but primers with two terminal C/G bases are recommended.
for increased PCR efficiency. Nucleotide residues C and G form a stronger pairing structure in the duplex DNA strands. Stability at the 3' end in primer template complexes will improve the polymerase efficiency. Sequence composition at the 5' end will be selected depending on the chosen task. The addition of bases not complementary to target sequences (‘tails’) is optional and can be added during the primer design procedure. The check for primer self-hybridization and cross-hybridization is performed simply; the number of nucleotide matches from the 3’ end is 6 with a single mismatch allowed to a target; for other parts of the primer 2 mismatches out of 12 bases are permitted. Mismatches at the 3’ end of primers are important as they decrease PCR efficiency.

We invented an abstract parameter called ‘primer quality’ that can help to estimate the efficiency of primers for PCR. ‘Primer quality’ is calculated by the consecutive summation of the points according to the following two parameters:

1. Primer order and purine-pyrimidine complexity, which is similar to linguistic complexity, measured by the alphabet-capacity l-gram method (Gabrielian and Bolshoy 1999; Orlov and Potapov 2004), self-complementarity and possible hairpin structures;
2. The melting temperature of the whole primer, the 12 terminal 3’ and 5’ and the central part of primers, determined primarily by the GC content.

This abstract value of ‘primer quality’ tries to describe the likelihood of PCR success of each primer; this value varies from 100% for the best to 0% for the “worst” primer. Although specificity is an important factor in PCR, primers may need to operate in a wider range of executable temperatures. For example, to meet multiplexing demands, it is possible to select the best primers with an optimal range of executable temperature using the program, allowing the design of qualified primers (probes) for any target sequence with any GC and repeat content. ‘Primer quality’ values of 80 and higher allow for the rapid choice of the best PCR primer pair combination. No adverse effects, due to the modification of the reaction buffer, sourced thermostable polymerases, or variations in annealing temperature, have been observed on the reproducibility of PCR amplification using FastPCR-designed primers.

A further advantage of FastPCR is its capability for designing primers from either degenerate or protein sequences, traditionally a difficult process, by utilising an auto-detection mechanism, based upon the standard nucleotide translation code, to overcome primer design limitations. For sequences with low GC content, users can choose special “low GC” options. By default, the primer design process does not allow overlapping primers, but in cases where no primers have been selected, the user can choose the “manual” option to list and check all variants of primers for their current location.

Two XLS files are generated by the program, showing the suggested primers and primer pairs in tabulated format for Microsoft Excel or Open Office. The spreadsheets display the following properties: automatically generated primer name, primer sequence, sequence location, direction, length, melting temperature, %GC content, molecular weight, molar extinction coefficient, linguistic complexity, and primer quality. For compatible primer pairs, the annealing temperature and PCR product size are also provided. In the file menu, users can learn the way to use the software for different tasks from several examples: primer design for multiplex PCR, degenerate DNA sequences, reading frame control, automatic detection of SSRs (simple sequence repeats) and design of primers around SSR regions, self-reporting primers (including LUX hairpin primer design) single primer PCR, and design of assays with one preset primer or to work with TaqMan® probes. The software can select primers for the individual tasks and parameters for each given sequence and also permit the linking up of several tasks to select a set of compatible primers – for example, detecting SSR loci in a set of sequences and designing primers for all of them.

**Hairpin (loop) and dimer formation**

Two types of primer-dimer, self-dimer and cross-dimer, may occur in a PCR reaction. FastPCR eliminates self and cross oligonucleotide reactions before generating a primer list and primer pair candidates. It is very important for PCR efficiency that the production of stable and inhibitory dimers are avoided, especially avoiding complementarity in the 3’-ends of primers from whence the polymerase will extend. Stable primer loop formation is very effective at inhibiting PCR since the loops formed reduce the ability of the primer to bind to the target; self complementarity means that two separate primers can interact in a similar way to create a thermodynamically more stable dimer (Fig. 2). The detection of primer dimers is based on analysis of non-gapped local alignments and the stability of the 3’-end and the central part of the primers. Primers will be rejected when they have the potential to form stable dimers with at least 6 bases at the 3’ end, containing no more than a single mismatch or two mismatches in 9 bases in the central part of the primer. These conditions are followed strictly for the control of dimers, and although there is a risk that some good primers will be eliminated, it is justified on the basis that reducing the cost of a reaction and expenditure of time are more important. FastPCR calculates the Tm for primer dimers with mismatches for pure and mixed bases using averaged nearest neighbour thermodynamic parameters provided for DNA/DNA duplexes.

**Calculation of optimal annealing temperature**

The optimal annealing temperature (Ta) is the range of temperatures where efficiency of PCR amplification is maximal without non-specific products. The most important values for estimating the optimal Ta is the primer quality and the Tm of the primers. Primers with high Tm’s (>60°C) can be used in PCRs with a wide Ta range compared to primers with a low Tm (<50°C). The optimal annealing temperature for PCR is calculated directly as the value for the primer with the lowest Tm. However, PCR can work in temperatures up to 10 degrees higher than the Tm of the primer, especially when reactions contain high primer concentrations (0.6-1.0 μM) to favour primer target duplex formation and increased Tm and Ta. For the short PCR fragments (<1000 bases) Ta values are equal to the Tm of the primers; for the same primer sets and longer PCR products (especially >3000 bases) we recommend increasing the Ta by one degree for every thousand base pairs. In our experience, almost all high-quality primers designed by FastPCR in the default or “best” mode guarantee amplification at annealing temperatures from 60 to 68°C without loss of PCR efficiency, and show good amplification in varying PCR annealing temperatures and when using different DNA polymerases and buffers.

![Fig. 2 Alternative structures formed by DNA molecules with inverted repeats. Intra-molecular interaction will give rise to hairpins (left), whilst inter-molecular hybridisation will give rise to ‘dimers’ (right).](image)
PCR amplification protocol

High quality primers help to perform PCRs even under sub-optimal PCR conditions. Because the primers have been designed to minimise primer-primer interactions, PCR reactions can be set up at room temperature and performed without hot-start enzymes. The range of optimal Ta’s is calculated using the Tm’s of the primers. Alternatively these values can be determined experimentally using gradient annealing temperature PCR, for example with a MasterCycler Gradient (Eppendorf). The highest temperature selected is usually about 10°C above the Tm of the primers. For primer combinations with very different Tm’s, the optimal annealing temperature is chosen according to the lowest primer Tm (primers with GC content higher than 50% are tolerant of a wide Ta range, from 50 up to 70°C).

In PCR reactions, for primer binding (usually from 55°C) and polymerase extension (usually from 55 to 72°C) steps, we recommend to join into one step at 60°C, thus performing two-step PCR. The recommended time for this step is one second for each 100 bp of PCR product. The denaturation of genomic DNA should be short to minimise inactivation of the enzyme at elevated temperatures; we recommend 95-96°C for 1-20 seconds. The PCR can be performed without hot-start enzymes. The range of optimal Ta’s is accessible at room temperature and performed without further heating. The most useful and popular program, Primer3 (NCBI/Primer-BLAST: http://www.ncbi.nlm.nih.gov/tools/primer-blast/), along with many other PCR primer design packages, do not perform this analysis for alternative amplification. In humans and animals, short interspersed nuclear elements (SINE), etc., are repeated many times (100 bp sequences) rather than using genomic DNA. Another possible way to solve this problem is to create a local specialized library of tandem and retrotransposons (especially SINE elements for animals and human genomes) repeat sequences: e.g., repeat libraries can be created from Repbase databases (available online: http://www.girinst.org/; Jurka et al. 2005).

By default, FastPCR performs a non-specific binding test of each given sequence. Additionally, the software allows this test to be performed against a reference sequence or sequences (e.g., BAC, YAC) or one’s own database. There is no necessity to scan primers through NCBI databases to prove specificity of PCR amplification. Preliminary detection that bind to more than one location on current sequences will be rejected. Even though the non-specific primer binding test is performed as a default for all primers, there is the option to cancel the function by the user.

Identification of secondary binding sites including mismatched hybridization is normally performed by considering the similarity of the primer to targets along the entire primer sequence. An implicit assumption is that stable hybridization of a primer with the template is a prerequisite for priming by DNA polymerase. However, DNA polymerases are known to be able to prime from an incomplete duplex formed between the primer and the template so long as the 3’ end of the strand elongation takes place is bound. FastPCR pays particular attention to the 3’ end portion of the primer (but not for probes) and calculates the similarity of the 3’ end of the primer to the target (length chosen by user) to determine the stability of the 3’-terminus. The reference non-specific primer binding test is based on a quick, non-gapped local alignment screening between the reference and input sequences.

The alignment generator for choosing the primer location on the reference sequence is performed by calculation, for both strands, using overlapping 7-mers from the 3’ end of the primer, of the local similarity for the 3’ end of the primer sequence. The conditions for quick alignment are changeable within certain limits: the minimum length is four bases for search initiation with at least 60% local similarity and a maximum of several mismatches at 3’ end of the primer sequence. This algorithm considers the various factors from one primer to another, depending on the properties of their sequences. The minimum length of 3’ end sequence matching the template required for priming varies from 8 to 15 bp.

For standard PCR conditions (55°C annealing temperature and 1.5 mM MgCl2) the primer can anneal to the target site and be extended by Taq polymerase if at least 12 bases from 3’ end of primers match the target, even given a single base mismatch at its 3’ end. This condition is optimal for...
performing quick alignment for searching for sequences complementary to the primer. For the search of the probe complement, the algorithm is performed without considering the 3‘ or 5‘ ends. The Tm for the primer-target duplex is calculated by the program.

Optionally, the user can synchronize the secondary non-specific primer binding test with a dataset of sequence names. The program recognizes that a given sequence in the secondary database (from loading the file) is the same name as the sequence for which it is designing primers and allows sequence selection to be made even though they match that screening sequence perfectly. This allows the same dataset to be used for both primer design and screening without having to make a new screening database for each sequence. In other words, a dataset that contains sequences A, B, C, and D can be used both for choosing primers and for checking primer specificity.

Single primer amplification, for targeting inverted repeats, is one of the most efficient PCR methods for detecting polymorphisms in genomes (Nelson et al. 1989; Sinnet et al. 1990; Welsh and McClelland 1990; Williams et al. 1990; Kalendar et al. 1999; Chang et al. 2001; Kalendar and Schulman 2006). FastPCR does not reject repeat regions on the target and reference sequences before searching for primer (probe) binding sites and it is therefore possible to design a primer (probe) to be specific to a repetitive region.

**Multiplex and degenerate primer design**

Multiplex PCR is an approach commonly used to amplify several DNA target regions in a single PCR reaction. The simultaneous amplification of many markers reduces the number of reactions that needs to be performed to amplify a set of markers; multiplex PCRs thus increase the throughput that can be performed with finite resources. The design of multiplex PCR assays can be difficult because it involves extensive computational analyses of primer pairs for cross interactions. FastPCR can quickly design a set of multiplex PCR assays for all the input sequences. PCR conditions may need to be adjusted; e.g., the annealing temperature increased or lowered so that all products are amplified equally efficiently. To accommodate this, most existing multiplex primer design packages use primer melting temperature. However, in practical terms, the design of almost identical Ta’s for each product is more important than identical Tm’s for each primer in the set. Also, the design of primers with identical Tm’s is difficult for different sequences. The melting temperatures of the PCR products are another important factor that is related to the annealing temperature value. The Tm of a PCR product depends on its GC content and length; short products are more efficiently amplified at low PCR annealing temperatures (100 bp - 50°C) than long products (3000 bp - >60°C). Further improvements can be achieved by selecting the optimal set of primers that maximize the range of common Tm’s. FastPCR quickly calculates multiplex primer pairs for the given target sequences and the number of target sequences and primer pairs for each of them.

An alternative way to design compatible multiplex PCR primer pairs is to use pre-designed primers as references for the design of new primers. The user can also select input options for the PCR products such as the minimum product size differences between the amplicons. This also allows the setting of primer design conditions individually for each given sequence. The speed of calculation depends on the settings of target sequences and primer pairs for each of them.

An alternative way to design compatible multiplex PCR primer pairs is to use pre-designed primers as references for the design of new primers. The user can also select input options for the PCR products such as the minimum product size differences between the amplicons. This also allows the setting of primer design conditions individually for each given sequence. The speed of calculation depends on the settings of target sequences and primer pairs for each of them.

**In silico PCR**

The hybridisation of primers to targeted annealing sites is only one way for PCR product prediction (Nishigaki et al. 2000; Lexa et al. 2001; Boutros and Okey 2004; Cao et al. 2005). Primers can bind to many predicted sequences in templates, but only attachment sites with few mismatches (one or two depending on their location) may permit the polymerase to perform strand elongation. The last 10-12 bases at the 3’ end of primers are important for general primer stability on the template and for the initiation of polymerase extension. Single mismatches in these last 10 bases at the 3’ end of primer can reduce the primer binding and PCR efficiency, the effect being stronger nearer to the 3’ end.

FastPCR allows simultaneous testing single primers or a list of the primer sequences designed for multiplex target sequences. It performs a fast, gapless alignment to test the complementarity of the primer to the target sequence. Parameters can be set to allow degrees of mismatches at the 3’ end of the primers. The program can also handle degenerate primers or primers/probes with 5’ or 3’ tail sequences.

Quick alignment for detection of primer locations on the reference sequence is performed by analysis on both strands of the DNA using overlapping n-mers (default = 7 bases) from the 3’ end of primer and calculating the local similarity of the primer to the target sequence. Parameters for quick alignment may be set: the minimum is 4 bases for search initiation and 60% local similarity. The probable PCR products can found for linear and circular templates using standard or inverted PCR, and for multiplex PCR. This in silico tool is very attractive for quickly analysing primers or probes against target sequences for determining primer location, orientation, and efficiency of binding, complementarity, and Tm.
Group-specific PCR primers (family-specific primer set) and unique PCR

Group-specific amplification, also called family-specific, sequence-specific, or universal amplification, is an important tool for comparative studies of genes and genomes that can be applied to studies of evolution, especially for gene families and for cloning new related sequences. The specific targets of these methods include disease resistance alleles (NBS profiling), transposable elements or other repeat elements (sequence-specific amplification polymorphism, S-SAP) that can be amplified to uncover DNA polymorphisms associated with these sequences. Once identified and characterised, FastPCR will help with PCR primer design around the polymorphic regions.

Group-specific primer sets are designed by first generating a multiple alignment of DNA sequences from family members and then manually identifying the most conserved regions for primer design. This can be slow, but is usually easy to understand. In comparison to the software ‘Primaclade’ (Gadberry et al., 2005), ‘Primique’ (Fredslund and Lange 2007) and ‘UniPrime’ (Bekaert and Teeling 2008), FastPCR does not use sequence alignment, giving it maximum flexibility to use a different strategy for PCR primer design. FastPCR does not design degenerate PCR primers to amplify a conserved (or polymorphic) region of all related sequences.

The overall strategy of designing group-specific PCR primers is divided into two phases: standard PCR design for the current sequence and then testing of complementarity of these primers to other sequences. The primer complementarity test is performed quickly with gapless local alignment, which includes parameters for setting the allowable number of mismatches at the 3’ end of the primers and for assessing primer similarity to the target sequence. Users can specify the alignment parameters (the same as for in silico PCR) for primers searching: “initial searching word size, >3 (default = 7), nt, important length at 3’ end, 5...20, nt for testing mismatches, minimum complement primer length (>12, nt) and the local similarity (default = 80%). The program generates an output page containing the group-specific PCR primers from each sequence and a second page showing compatible primer combinations with details of product sizes and suggested annealing temperature for PCR.

FastPCR designs large sets of universal primer pairs for all given sequences, identifies conserved regions, and generates suitable primers for all given sequences. All steps of the algorithm are performed automatically and the user can influence the general options for primer design and alignment. The software has been experimentally tested for very different tasks and sequences. FastPCR will work with any source of sequences as long as it is possible to find short consensus sequences among the sets. The quality of primer design is dependent on sequence relationships, phylogenetic similarity, and suitability of the consensus sequence for the design of good primers. The software is able to generate group-specific primers for each set of sequences independently, which are suitable for all sequences. The strategy for designing a unique PCR primer is different from designing group-specific PCR primers. In this case, the program searches for unique regions within a DNA sequence and automatically designs primers with minimal user intervention and maximum flexibility. Primer alignment parameters for group-specific PCR primers are similar to those used for in silico PCR.

Real-time PCR primers and probe design

Real-time PCR (RT-PCR) is a highly sensitive method that can be used for the detection and quantification of target sequences without gel electrophoresis. Generally, RT-PCR primer design is not different from standard or multiplex PCR primer design. However, one major difference is that, in the absence of electrophoresis to separate and visualise the products, there is no need to design differently sized amplicons. Moreover, it is important that amplicons that are multiplexed for quantification are close in size to mitigate any effect size difference may have on PCR efficiency. The TaqMan® assay is one of the most widely used methods for DNA quantification by RT-PCR (Yoshimura et al. 2005). It uses a primer pair to amplify the target region. A third oligonucleotide acts as a probe for the target region. This oligonucleotide is dual-labeled with a fluorophore/reporter residue quencher-labeled peptide nucleic acid (Q-PNA) probe and two primers. The Q-PNA hybridizes to a complementary tag sequence located at the 5’ end of a 5’ fluorophore-labeled oligonucleotide primer, quenching the primer’s fluorescence. A 13-residue quencher-labeled peptide nucleic acid (Q-PNA) bearing a C-terminal DABCYL group is hybridized to various 5’ fluorophore-labeled oligonucleotides that contain a common 5’ Q-PNA complementary region and 3’ target-specific primer sequences. The length of the Q-PNA is chosen so that the Q-PNA/primer duplex has a Tm higher than the primer annealing temperature and lower than the Tm of the primer/amplicon duplex. As a result, excess primers are quenched at the annealing temperature and the fluorescence is measured during the annealing step as an indication of the amount of primer hybridized to amplicon i.e., the amount of product.

Molecular beacon assays, like TaqMan®, use a single oligonucleotide probe with both reporter and quencher but at either end. Inverted repeats at their ends allow the molecules to form hairpins in solution (see Fig. 2) bringing the reporter and quencher together so the fluorophore is naturally quenched. The loop portion of the molecule is a probe sequence complementary to the amplified nucleic acid target. Fluorescence changes due to product accumulation through probe degradation and/or hybridisation are measured. Molecular beacons are useful in situations where it is either not possible or desirable to isolate the probe-target hybrids from an excess of the hybridization probes, such as in real time monitoring during PCR in sealed tubes or in detection of RNAs within living cells.

Another self-reporting molecule method is LUX (light upon extension) RT-PCR (Nazarenko et al. 2002). The use of primers for detecting and quantifying genes in real-time platforms represent the most recent advance in real-time chemistry design. LUX differs from TaqMan® in that:

- the reporter is sited on the PCR amplification primer and not located on a self-quenching probe;
- fluorescence quenching is achieved through the ‘natural’ hairpin structure of the labeled primer without the need for a quencher; and
- release of quenching is caused by the loss of the hairpin structure achieved through DNA synthesis and not probe degradation or hybridisation.

The design of LUX primers is based on studies that demonstrate the effects of the primary and secondary structures of oligonucleotides on the emission properties of a conjugated fluorophore. Design factors are largely derived from the necessity of having guanosine bases in the primary sequence near the conjugated fluorophore. LUX primers utilize one single-labeled fluorogenic primer and a corresponding unlabeled primer. No probes or quenchers are needed. A hairpin structure provides fluorescence quenching of the fluorophore. When the primer is incorporated into dsDNA, the fluorophore is dequenched, significantly increasing the

Real-time PCR primers and probe design

Real-time PCR (RT-PCR) is a highly sensitive method that can be used for the detection and quantification of target sequences without gel electrophoresis. Generally, RT-PCR primer design is not different from standard or multiplex PCR primer design. However, one major difference is that, in the absence of electrophoresis to separate and visualise the products, there is no need to design differently sized amplicons. Moreover, it is important that amplicons that are multiplexed for quantification are close in size to mitigate any effect size difference may have on PCR efficiency. The TaqMan® assay is one of the most widely used methods for DNA quantification by RT-PCR (Yoshimura et al. 2005). It uses a primer pair to amplify the target region. A third oligonucleotide acts as a probe for the target region. This oligonucleotide is dual-labeled with a fluorophore/reporter residue quencher-labeled peptide nucleic acid (Q-PNA) probe and two primers. The Q-PNA hybridizes to a complementary tag sequence located at the 5’ end of a 5’ fluorophore-labeled oligonucleotide primer, quenching the primer’s fluorescence. A 13-residue quencher-labeled peptide nucleic acid (Q-PNA) bearing a C-terminal DABCYL group is hybridized to various 5’ fluorophore-labeled oligonucleotides that contain a common 5’ Q-PNA complementary region and 3’ target-specific primer sequences. The length of the Q-PNA is chosen so that the Q-PNA/primer duplex has a Tm higher than the primer annealing temperature and lower than the Tm of the primer/amplicon duplex. As a result, excess primers are quenched at the annealing temperature and the fluorescence is measured during the annealing step as an indication of the amount of primer hybridized to amplicon i.e., the amount of product.

Molecular beacon assays, like TaqMan®, use a single oligonucleotide probe with both reporter and quencher but at either end. Inverted repeats at their ends allow the molecules to form hairpins in solution (see Fig. 2) bringing the reporter and quencher together so the fluorophore is naturally quenched. The loop portion of the molecule is a probe sequence complementary to the amplified nucleic acid target. Fluorescence changes due to product accumulation through probe degradation and/or hybridisation are measured. Molecular beacons are useful in situations where it is either not possible or desirable to isolate the probe-target hybrids from an excess of the hybridization probes, such as in real time monitoring during PCR in sealed tubes or in detection of RNAs within living cells.

Another self-reporting molecule method is LUX (light upon extension) RT-PCR (Nazarenko et al. 2002). The use of primers for detecting and quantifying genes in real-time platforms represent the most recent advance in real-time chemistry design. LUX differs from TaqMan® in that:

- the reporter is sited on the PCR amplification primer and not located on a self-quenching probe;
- fluorescence quenching is achieved through the ‘natural’ hairpin structure of the labeled primer without the need for a quencher; and
- release of quenching is caused by the loss of the hairpin structure achieved through DNA synthesis and not probe degradation or hybridisation.

The design of LUX primers is based on studies that demonstrate the effects of the primary and secondary structures of oligonucleotides on the emission properties of a conjugated fluorophore. Design factors are largely derived from the necessity of having guanosine bases in the primary sequence near the conjugated fluorophore. LUX primers utilize one single-labeled fluorogenic primer and a corresponding unlabeled primer. No probes or quenchers are needed. A hairpin structure provides fluorescence quenching of the fluorophore. When the primer is incorporated into dsDNA, the fluorophore is dequenched, significantly increasing the
fluorescent LUX signal. Because such primers are at an early stage of development, little literature on them is available. They do, however, appear to represent a more economical option than other detection chemistries and are relatively easy to design. Claimed advantages of LUX are that dual-labeled probes are technically difficult to design, more expensive to manufacture and are more prone to (thermal) degradation leading to higher background noise with a loss in sensitivity.

**Primer analysis**

Individual or sets of primers are evaluated by FastPCR for the following: calculated Tm using default or other formulae for normal and degenerate nucleotide combinations; GC content; extinction coefficient; unit conversion (nmol per OD); mass (μg per OD); molecular weight; linguistic complexity; primer quality. All primers are analyzed for intra- and inter-molecular interactions to form dimers. The web version for on-line primer testing with Tm calculations for normal and degenerate nucleotide combinations based on nearest neighbour thermodynamic parameters is available at [http://primerdigital.com/Tools/PrimerAnalyser.html](http://primerdigital.com/Tools/PrimerAnalyser.html). Comprehensive analyses of the list of primers with prediction of the oligonucleotide properties, self and cross dimer detection and temperature of annealing calculation are located at [http://primerdigital.com/Tools/PrimerList.html](http://primerdigital.com/Tools/PrimerList.html).

**Sequence manipulation tools**

The FastPCR program includes tools for the manipulation and transformation of DNA and protein sequences. Sequences can be reformatted and shown as the complement, reverse, or antisense sequence, or translated to encoded polypeptides in all reading frames. The output file can be in FASTA format or exported to Excel. FastPCR also has a special tool for screening for vector contamination, which includes a list of standard primers. The software scans the sequences to remove contaminating primer sequences derived from commonly-used vectors based on a local sequence alignment. The primers can be sorted (clustered) with similarity to each other using k-mer counting (a k-mer is a contiguous subsequence of length k). FastPCR allows the alignment search of a DNA sequence or a sequence list within personal databases (similar to BLAST) (Altschul et al. 1990; Morgenstern 1999). This is convenient when selecting primers and for in silico monitoring.

**Alignment algorithm and repeat search**

Repeat detection and efficient alignment are well-studied problems in bioinformatics. A multitude of computational tools have been developed for repeat detection using automated multiple alignment programs based mostly on sequence analysis. There are many papers describing algorithms for finding repeats in a string; in most cases, alignment methods for finding repeats start with a self-comparison to detect repeated sequences (Kurtz 2000; Kurtz et al. 2001; Bao and Eddy 2002; Lefebvre et al. 2003; Campagna et al. 2005; Edgar and Myers 2005).

FastPCR implements a method for discovering repeated sequences within genomes using only the genomic sequence itself. Global repeats are families of related (or nearly identical) sequences that occur multiple times. The FastPCR method for identifying repeats is based on defining high-frequency k-mers (short substrings of length k) as seeds, and then extending each seed to a progressively longer sequence, followed by clustering for consensus sequence alignment. Following this process, the software is capable of detecting repeats, even degenerate repeats. To find all degenerate repeats (or substrings) the user specifies alignment parameters: initial k-mer size, minimum extended substring length with minimum similarity, maximum gap size and minimum repeat length. As a basis for alignment of sequences the following rules are followed: related sequences should have sequential common homologous blocks (>65% similarity); inversions, deletions or permutations are looked upon as gaps in sequence comparisons.

The alignment algorithm starts searching for an exact k-mer (default = 7, minimum 2 bases) sequence which it then extends to each side of the original sequence to find the best alignment, without a gap, that fulfills the set similarity parameters. The extension is done by pairwise character comparison. The algorithm directly computes the maximum exact segment substrings for many candidates. Finally, the segment substrings are joined to form a long string with gaps. For detection using highly degenerate substrings, FastPCR allows the application of purine-pyrimidine local alignment. In aligning the DNA sequences the program uses a similarity table for all degenerate nucleotides, which increases the probability of identifying related sequences. K-mer substrings are compressed using multidimensional and jagged arrays. For example, with a k-mer is of 7 and the 4 nucleotide code (A-0, G-1, C-2, G-3), we create 7 multidimensional arrays, for 7 positions and a jagged dimension used for amount [0] and k-mers location coordinates: int[[], , , , , ,][] X = new int[4, 4, 4, 4, 4, 4, 4][];

This array is very compact in memory and convenient for alignment. The software uses 7- or 11-mers for quick alignment and any k-mer in a standard alignment algorithm. This fast and effective DNA analysis algorithm has been developed for the search of composite and long repeats by alignment in the following formats: direct - direct, direct - inverted, direct - antisense, direct - inverted antisense. Searching in all orientations facilitates the rapid identification long direct repeats, transposons, tandem repeats and hairpin repeats (short inverted repeats). FastPCR provides an easy-to-use interface for examining repeat type and its structure. The output of the search is displayed in a diagram showing the repeat location for each pair as blue and red lines to depict direct and inverted repeats, respectively, and as text files with alignment. For example, in Fig. 3, the FastPCR output display shows the repeat searching result for the BAC Triticum monococcum genomic sequence (AY485644; Yan et al. 2002). The long terminal repeats (LTR retrotransposons) are shown as blue lines in the position 109,708 and 153,591 nt. The other repeats are complex and contain several overlapping direct repeats within a short region.

The software can be applied to reveal the repetitive sequences of chromosomes or genomes (Fig. 4), which is useful because eukaryotic genomes have a high content of repetitive DNA, both tandem and dispersed. Fig. 4 shows that 26.5% of the 30.4 Mb sequence of Arabidopsis thaliana chromosome 1 is covered by reversed complements of the centromere of chromosome 1 is easy to detect in the middle of picture by the special structure of the clusters of centromeric repeats.

An extension of the ability of FastPCR to align sequences makes it able to perform repeat searches without visualisation of inter-sequence comparisons for given sequences as a “Clustering” tool. The algorithm for searching direct repeats is based on the local alignment of a current sequence against another sequence. For other kinds of repeats such as inverted repeats, local alignment is performed using the reverse-complement sequence, which would identify some classes of transposable elements that have inverted repeats such as ‘foldback’ elements and MITEs. Similarly, two other unusual variants for repeat searching can be performed: original sequence against a reverse sequence and original sequence against a complement sequence.

**An efficient tool for discovering LTR retrotransposons**

LTR retrotransposons exist in virtually all eukaryotic genomes and have been important in shaping genomes, especially of plants, during evolution. They comprise 50% or more of many cereal genomes, which shows their importance in determining chromosome structure, function, and
Genes, Genomes and Genomics 3 (Special Issue 1), 1-14 ©2009 Global Science Books

Overall genome dynamics. Their similarity to retroviruses, in terms of structure and replication mechanism, suggests a common origin. Similar retrotransposons have been found both between species and within genomes (Flavell et al. 1992), both in related and in diverse species (Kalendar et al. 2004, 2008), making them biologically very interesting. As molecular markers, no other family of markers gives the genome coverage and polymorphic information content of retrotransposons. The need for sensitive and accurate detection and annotation of retrotransposons grows commensurately with the speed of genome sequencing.

Retrotransposons are typically characterized by two identical long terminal repeat (LTR) sequences flanking the coding region for genes necessary for reverse transcription of the RNA and integration into the genome. The internal region contains the following functional domains: primer binding site (PBS), which is a region complementary to a tRNA 3’/5’ terminal sequence used during reverse transcription; the Gag region, coding a capsid protein; the Pol region, coding for protease, integrase, and reverse transcriptase enzymes; additionally for some LTR retrotransposons the Env region, which contains the gene coding for an envelope protein; at the 3’ end, a purine-rich sequence, called the poly-purine tract (PPT). Some groups of LTR retrotransposons, such as TRIMs and LARDs, have very short internal regions incapable of encoding any proteins; these retrotrans-

Fig. 3 Results of repeat search using FastPCR. The output shows the result of repeat searching in BAC 438,828 bp Triticum monococcum genomic sequence (AY485644; Yan et al. 2002) and presented in a graphical display demonstrating that 65% of the genomic sequence is covered with repeated sequences). Repeat families are represented by the Y axis: the presence (with position along the BAC clone) and size of a repeated sequence is represented by the X-axis.

Fig. 4 Show that 26.5% from 30.4M bp sequence of Arabidopsis thaliana chromosome 1 (available online: ftp://ftp.arabidopsis.org/home/tair/Sequences/) covering by repeat sequences. The centromere of chromosome 1 is easy to detect in the middle of picture by the special structure of the clusters of centromeric repeats.
posons have only PBS and PPT noncoding sequences between the LTRs, presumably parasitizing the enzymes required for propagation (replication and transposition) from other active elements.

There is already some software specifically developed for identifying LTR retrotransposons – ‘LTR STRUCT’ (McCarthy and McDonald 2003), ‘LTR_FINDER’ (Xu and Wang 2007), ‘LTR_par’ (Kalyanaraman and Aluru 2006), LTR_Rho (Wang 2007), ‘LTR_par’ (Kalyanaraman and Aluru 2006), LTR_Rho (Wang 2007), and ‘LTRharvest’ (Ellinghaus et al 2007). The PBS is complementary to the 3′ termini of cellular tRNAs, which act as primers for cDNA synthesis by reverse transcriptase. In retroviruses, the PBS is complementary to 18 nt at the 3′ terminus of the primer tRNA, and to 8–18 nt in retroposons (Marquet et al 1995; Mak and Kleiman 1997; Kelly et al 2003; LeGrice 2003). The PBS of LTR retrotransposons and all retrovirus PBS are complementary to one of a limited set of tRNAs: tRNA^{ Met}_{ Met}, tRNA^{ Lys}_{ Lys}, tRNA^{ Pro}_{ Pro}, tRNA^{ Trp}_{ Trp}, tRNA^{ Gln}_{ Gln}, tRNA^{ Arg}_{ Arg}, tRNA^{ Phe}_{ Phe}, tRNA^{ Lys}_{ Lys}, or tRNA^{ Gln}_{ Gln}. The tRNA choice appears to depend on the group and one tRNA may predominate: lentiviruses, including HIV-1, use tRNA^{ Met}_{ Met} (Mak and Kleiman 1997), whereas tRNA^{ Met}_{ Met} is commonly, though not exclusively, used by plant retrotransposons. A handful of exceptional LTR retrotransposons, including the Tfl/sushi group of fungi and vertebrates and Fourf in maize, self-prime cDNA synthesis rather than use a tRNA primer (LeGrice 2003; Hizi 2008).

The LTR retrotransposons are generally identified by similarity to diagnostic protein coding domains, particularly reverse transcriptase and integrase or to LTRs of other known retrotransposons. However, protein coding domains are absent from LARD and TRIM retrotransposons (Witte et al 2001; Kalander et al 2004, 2008). Structure-based identification of LTRs, by the presence of long direct repeats, is the only viable approach for identifying these elements. The presence of a PBS adjacent to a direct repeat is a strong signature for a retrotransposon. We found that we could identify new LTR retrotransposons without relying on a library of known LTRs or coding sequences by searching for PBS sequences. We used the PBS to identify new elements and to place them in both superfamilies of LTR retrotransposons (Gypsy and Copia) as well as the non-autonomous LARD and TRIM elements.

LTR retrotransposon identification follows a number of steps: Firstly, for general repeat detection, the algorithm builds a set of repeat clusters by using exact k-mers (10 bases, for example) as seeds, and extends each seed to a progressively longer consensus sequence (allowing errors, but not gaps). The most frequent k-mers are clustered for identifying longer repeats and the program then annotates regions with these repeat clusters. The minimal size of a repeat cluster helps to identify and define the 3′-end of the left LTR. If the sequence near the PBS site is found to be repeated and annotated (see first step), this is taken as further evidence that the sequence is probably the 3′-LTR of a retrotransposon. Together the two steps identify reliable candidates for LTRs.

The termini of LTRs are not always bound by the motif 5′-TG…CA-3′, but can be 5′-TA…CG-3′ (in general: 5′-TR…CR-3′). The sequences of LTR termini are not always perfect, mutated to give mismatches, insertions and deletions and, as such, not a reliable indicator of LTR sequences. For correct LTR identification several (at least two) complete LTRs are required. All partial or complete LTR sequences are aligned to produce a consensus sequence. In reality, perfect LTR retrotransposons are rare. Many retrotransposons have undergone dramatic changes, often through the transposition of others within each other to form a nested, “Russian Doll” structure, followed by subsequent deletions and LTR-LTR recombinations. The consequence of this is a puzzle of retrotransposon sequences where LTRs, and internal sequences from a number of retrotransposons, are mixed in an apparently labyrinth manner. Many LTR retrotransposons are old and not intact; their LTR sequences are not easily recognizable using existing software. Protein coding domains are absent from LARD and TRIM retrotransposons, so it is not possible to identify these by common protein domains. All these problems bring difficulties and may result in the wrong interpretation or a putative LTR retrotransposon. The solution is partly manual analysis of each individual LTR retrotransposon, starting with linkages between LTRs of a given family, the consensus sequence(s) can be deduced by sequence alignment, from which it is possible to analyse the internal domains, if they (still) exist, between these LTRs.

The steps described for LTR retrotransposon discovery were implemented in software, which we tested by applying our program to find LTR retrotransposons in several plant and animal species (Benjak et al 2008; Kalander et al 2009). We report in silico computational methods that can identify new LTR retrotransposons without relying on a library of known elements and reverse transcriptase sequences. Our in silico-search approach proved to be very effective: primer binding sites (PBS), were identified by similarity searches against a local database of PBS sequences. We screened 149 PBS sequences with 12 bases allowing 2 mismatches. The internal domain of the identified retrotransposons started with PBS sequences (TGG) located from 0 to 10 nt after the LTRs, and finished with PPTs, allowing us to extract the LTRs.

This method identifies LTRs and complete retrotransposons by using PBS sequence searching and LTR clustering. Our method can identify in silico all types of LTR retrotransposons (e.g., Gypsy, Copia, LARD, and TRIM elements). We tested this on plant and animal genomes: Oryza sativa, Vitis vinifera, Arabidopsis thaliana, Solanum lycopersicum and Drosophila melanogaster. We identified more known LTR retrotransposons than have been found before with other methods. Specifically, our method identifies any LTR retrotransposons which are present as a single complete element or truncated with LTR and internal part, in chromosome sequence. We found many new putative retrotransposons, even rice ones, which currently are not annotated and absent from the Oryza sativa RefOrzya database. Likewise, the searches in Solanum lycopersicum and Vitis vinifera databases produced sequences for numerous new retrotransposon elements. Some examples and accession numbers for the sequences resulting from this study have been published (Benjak et al 2008; Kalander et al 2009).

Simple sequence repeat (SSR) loci search

Simple sequence repeats (SSRs, or microsatellites) are tandemly repeated sequences with one or more bases as a core. Microsatellites are ubiquitously distributed throughout eukaryotic genomes and many SSRs are highly polymorphic in length, arising from changes in the number of repeats. Variations in SSR sequences originate from the specific DNA structure of SSR regions (such as Z-DNA for purine-pyrimidine simple repeats: (CT)n or (AG)n, pers. comm.) that...
affects DNA (or RNA) polymerase during replication.

SSR loci are an important class of markers for population genetic studies despite their development requiring screening of genomic DNA libraries, sequencing of clones, designing of primers, testing, and validating the assays. The common practice of size-fractionating genomic DNA before cloning could lead to differential representation of SSR loci within genomic libraries. In addition, linkage mapping studies have shown that small numbers of SSR markers are not randomly distributed within the genomes from which they are isolated. However, the increased availability of DNA sequences in the databases has allowed the identification of SSRs using an in silico approach, either from genomic or cDNA sequences (Abajian 1994; Jewell et al. 2006; da Maia et al. 2008).

Our approach to SSR searching is based on analysing the low-complexity regions with the quick linguistic sequence complexity method. This method allows the detection of perfect and imperfect SSRs with a single, up to 10-base, core-tandem motif. Each entry sequence is processed for identification of SSRs and the flanking DNA sequence is used for design of compatible forward and reverse primers for their amplification by PCR. FastPCR is capable of automatically identifying all SSRs within each entry sequence and individually designing compatible PCR primer pairs for each SSR locus. The default PCR primer design parameters are that the primers must be within 100 bases from either side of the identified SSR. Often the sequence available around SSR loci is not sufficient for designing good primers and the user can increase or decrease the distance from either side to find more efficient and compatible primer pairs. The FastPCR program has been successfully implemented for designing PCR primers of SSR loci in many different laboratories throughout the world and some articles are shown on our web page. The capabilities of FastPCR make it a complete bioinformatics tool for the use of microsatellites as markers, from discovery through to primer design.

DISCUSSION

We have developed an efficient and adapted-for-practice package for PCR primer and probe design as well as for searches for repetitive sequences. The software can work simultaneously with multiple nucleic acid or protein sequences. This free bioinformatics tool, FastPCR, was developed, and continues to be improved, based on detailed experimental studies of PCR efficiency for the optimal design of primer and probe sequences.

FastPCR is a toolbox for many functions; it can: calculate the thermal properties of DNA sequences and, applying this information, design primers for use in standard PCR, long distance PCR, inverse PCR, real-time LUX, multiplex, group-specific (common primers for given N target sequences), and unique (design of specific PCR primers for each sequence) PCR, as well as for single primer (primers from closely located inverted repeats) applications; automatically detect SSR loci and direct PCR primer design; design degenerate primers from amino acid sequences; design long oligonucleotides for microarray analyses and dual-labeled oligonucleotide probes as molecular beacons; carry out "in silico" PCR and probe searches for discovering oligonucleotide target binding sites and for melting temperature calculation, comprehensive primer pairs and individual primer analysis tests included; perform sequence alignments to identify repetitive DNA sequences that are highly abundant in the genomes of higher organisms.

We have used our own findings and data from other laboratories to predict more precisely oligonucleotide “quality”. Using these to set the parameters for oligonucleotide design we were able to improve our success rate for PCR. In contrast to other PCR primer design software, FastPCR was tested on large data sets including very different eukaryotic and prokaryotic genomes for the very different PCR tasks as listed above.

FastPCR both predicts primers and oligomers of high quality and provides tools for visualization and management of large and routine (everyday) projects and also proofs and testing of long oligonucleotides. The analysis of the quality of potential primers is performed differently from other software packages because only a few basic parameters (thermodynamic stability, stability of secondary binding sites and dimer formation) are evaluated. Properties such as GC content, dimer formation, specific bases at the 3’ terminus of primers, purine-pyrimidine tracks, sequence linguistic complexity, and secondary (non-specific) binding are analyzed and used for generating the best primer combinations. For generating PCR primers, FastPCR can be used, not only for gene-specific targets, but also for repetitive DNA. The software also supports the design of primers matching a specific reading-frame of a coding region, thus facilitating the application of the resulting PCR fragments in expression cloning.

FastPCR is flexible in addressing further applications. The design of primers for the group-specific amplification from related genomes or of sequence from the same genome is easy with FastPCR. For example, a group-specific primer related to storage proteins in mature cereal grains (prolamin and globulin) can be designed. These proteins are highly conserved but the DNA sequences of the members are polymorphic within and between species. Applications in single nucleotide polymorphism (SNP) detection are possible.

CONCLUSION

FastPCR has helped design many thousands of PCR primers and many probes (e.g., TaqMan®) for a range of very different applications and tasks in our laboratory, which have given excellent performance in PCR amplification. The majority of primers were successfully used in the different applications described above. The software algorithms can select highly specific oligonucleotides suitable for any foreseeable PCR application, with a wide functional temperature range, and eliminate potentially problematic oligonucleotides using secondary structure prediction. This software is flexible and allows the application of its modular functionality to wide range of projects.

The FastPCR program has been available on the internet since 1999 and has been tested in numerous laboratories and companies around the world. Its extensive trialling has provided valuable feedback that has been used to develop the program further. We receive a lot recommendations and suggestions for the improvement of the software design and programming, together with criticisms and error reports. These are important for further development and improvements in response to the needs and innovative ideas of its users. We will continue to evaluate FastPCR privately and publicly through the web of labs collaborating with us, as well as world-wide voluntary researchers providing feedback.

AVAILABILITY

FastPCR is available free to academic institutions, provided that it is used for non-commercial research and education only. It is not to be reproduced or distributed for commercial use. The program manual, licence agreement, and files for installation are available on the internet at http://www. biocenter.helsinki.fi/bi/Programs/. Web tools include: java applets for comprehensive on-line primer analysis and melting temperature calculations for normal and degenerate nucleotide combinations based on the nearest neighbour thermodynamic parameters; comprehensive oligonucleotide dilution and resuspension calculator; applet for comprehen-
sive analysis of the list of primers with prediction of the oligonucleotide properties and self and cross dimer detection; annealing temperature calculation and general PCR and qPCR set-up applet. These are available here: http://primerdigital.com/Tools/. A web interface (Java) and stand-alone software (VS.Net C#) for Linux and Mac operation systems are currently under development.

ACKNOWLEDGEMENTS

Supported by the ERA-PG project EXBARDIV.

REFERENCES

Allawi HT, Sambandam J (1997) Thermodynamics and NMR of internal G-T mismatches in DNA. Biochemistry 36, 10581-10594
Novire L (2001) MELTING, a fast tool to compute the melting temperature of nucleic acid duplexes. Bioinformatics 17, 1226-1227

PCR primer and probe design and repeat research. Kalendar et al [48x373]
Williams JGK, Kubelik AR, Livak KL, Rafalski JA, Tingly SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18, 6513-6535
Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proceedings of the National Academy of Sciences USA 98, 13778-83
Yang X, Scheffler BE, Weston LA (2006) Recent developments in primer design for DNA polymorphism and mRNA profiling in higher plants. Plant Methods 2, 4