Genes, Genomes and Genomics *©2009 Global Science Books*

FastPCR Software for PCR Primer and Probe Design and Repeat Search

Ruslan Kalendar[1,2*] • David Lee[3] • Alan H. Schulman[1,4]

[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-31600 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

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.

Received: 13 February, 2009. Accepted: 11 July, 2009

Invited Review

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.htm) 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 executive 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, nmole and mg per OD_{260}), 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; polypyrimidine (T, C) or polypurine (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.

Criteria	Default	Ideal
Length (nt)	20 – 25	22 – 35
Tm range (°C)[a]	55 – 65	60 – 68
Tm[a] at 3′-end, 12 bases	30 – 48	41 – 45
Tm[a] at 5′-end, 12 bases	30 – 45	38 – 45
GC (%)	45 – 65	50
3′-end composition (5′-nnn-3′)	sws, wss, ssw, swa	wss, ssa, sws
5′-end composition (5′-nnn-3′)	nnn	tsn
Sequence linguistic complexity[b]	>80%	>90%
Purine-pyrimidine linguistic complexity[b]	>80%	>90%
Degenerate bases at 3′ end, 12 bases	< 3	0
‘Primer Quality’	80	>90

[a] Nearest neighbour thermodynamic parameters (Allawi and SantaLucia 1997).

[b] Sequence linguistic complexity measurement was performed using the alphabet-capacity l-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

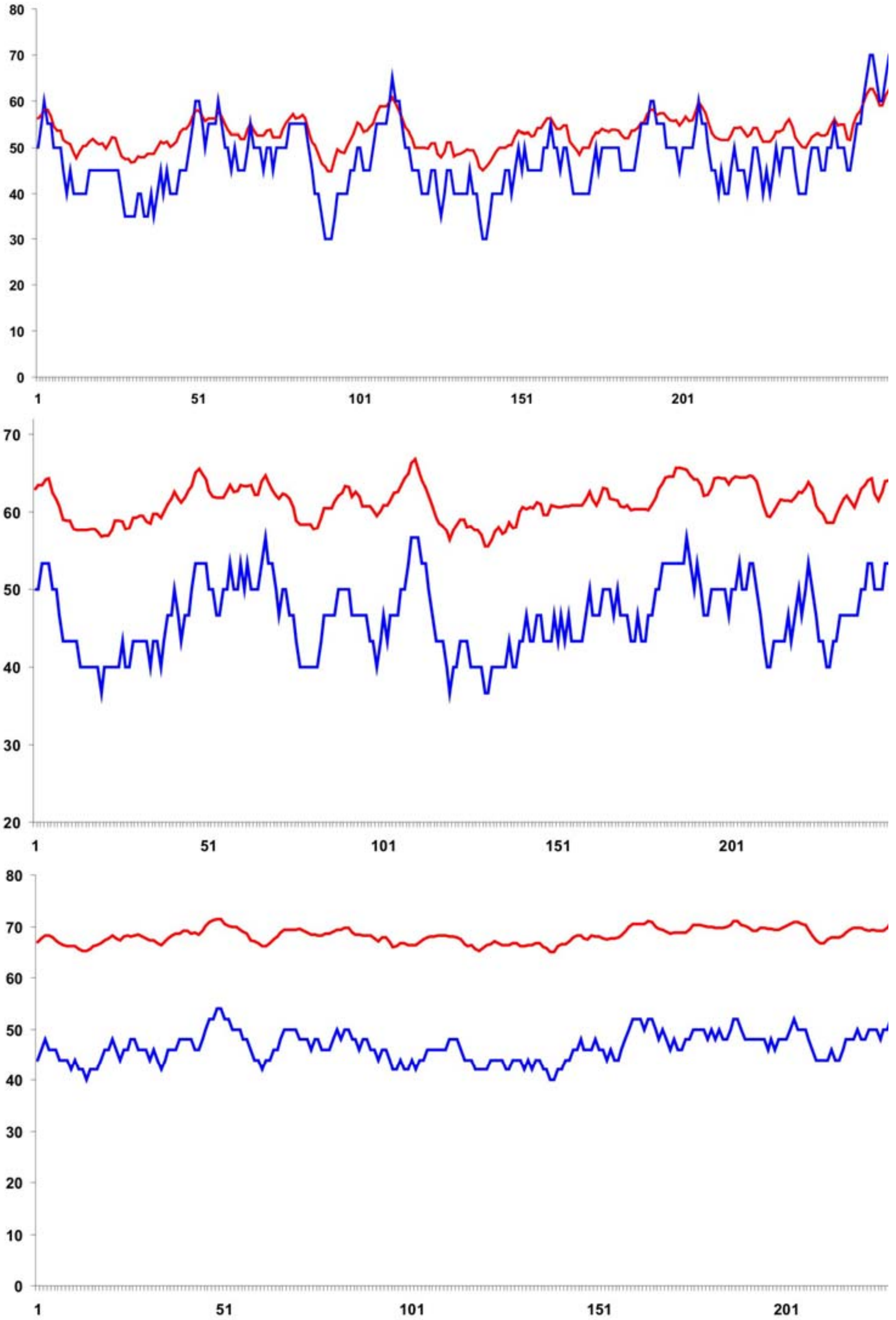

Fig. 1 Effect of GC content on the melting temperature of DNA. The melting temperature of a stretch of DNA is shown. The blue lines represent the %GC content of the sequence with the calculated melting temperatures (red line in °C) using the nearest neighbour thermodynamic parameters of SantaLucia (1998). The top, middle and bottom panels represent DNA window sizes of 20, 30, and 50 bases respectively.

"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; Bommarito *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 sequences 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 coefficient is similarly predicted by averaging nearest neighbour values at mixed sites (**Tables 2, 3**). Similar averaging by Allawi's nearest neighbour thermodynamic parameters for melting temperature calculation is used in the new web tool - IDT SciTools (Integrated DNA Technologies, Owczarzy *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 $5'\mathbf{N_1}N_2$ is shown in the horizontal column and the second nucleotide $5'N_1\mathbf{N_2}$ 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 formulae: 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 14 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 (Gabrielian and Bolshoy 1999; Orlov and Potapov 2004) within 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

Table 2 Unified oligonucleotide dH^0 (cal mol^{-1}) for normal and mixed nucleotides. Horizontal nucleotides for first nucleotide $5'\mathbf{N_1}N_2$ and vertical column for second nucleotide $5'N_1\mathbf{N_2}$.

5′-3′	**A**	**B**	**C**	**D**	**G**	**H**	**K**	**M**	**N**	**R**	**S**	**T**	**V**	**W**	**X**	**Y**
A	-7600.0	-7800.0	-8400.0	-7533.3	-7800.0	-7733.3	-7500.0	-8000.0	-7750.0	-7700.0	-8100.0	-7200.0	-7933.3	-7400.0	-7750.0	-7800.0
B	-7966.7	-8577.8	-8666.7	-8344.4	-9033.3	-8222.2	-8533.3	-8316.7	-8425.0	-8500.0	-8850.0	-8033.3	-8555.6	-8000.0	-8425.0	-8350.0
C	-8500.0	-8800.0	-8000.0	-8966.7	-10600.0	-8100.0	-9200.0	-8250.0	-8725.0	-9550.0	-9300.0	-7800.0	-9033.3	-8150.0	-8725.0	-7900.0
D	-7666.7	-8244.4	-8800.0	-7866.7	-8100.0	-8100.0	-7966.7	-8233.3	-8100.0	-7883.3	-8450.0	-7833.3	-8188.9	-7750.0	-8100.0	-8316.7
G	-8200.0	-8733.3	-9800.0	-8200.0	-8000.0	-8800.0	-8200.0	-9000.0	-8600.0	-8100.0	-8900.0	-8400.0	-8666.7	-8300.0	-8600.0	-9100.0
H	-7766.7	-8266.7	-8200.0	-8122.2	-8966.7	-7866.7	-8300.0	-7983.3	-8141.7	-8366.7	-8583.3	-7633.3	-8311.1	-7700.0	-8141.7	-7916.7
K	-7700.0	-8466.7	-9000.0	-8033.3	-8250.0	-8283.3	-8200.0	-8350.0	-8275.0	-7975.0	-8625.0	-8150.0	-8316.7	-7925.0	-8275.0	-8575.0
M	-8050.0	-8300.0	-8200.0	-8250.0	-9200.0	-7916.7	-8350.0	-8125.0	-8237.5	-8625.0	-8700.0	-7500.0	-8483.3	-7775.0	-8237.5	-7850.0
N	-7875.0	-8383.3	-8600.0	-8141.7	-8725.0	-8100.0	-8275.0	-8237.5	-8256.3	-8300.0	-8662.5	-7825.0	-8400.0	-7850.0	-8256.3	-8212.5
R	-7900.0	-8266.7	-9100.0	-7866.7	-7900.0	-8266.7	-7850.0	-8500.0	-8175.0	-7900.0	-8500.0	-7800.0	-8300.0	-7850.0	-8175.0	-8450.0
S	-8350.0	-8766.7	-8900.0	-8583.3	-9300.0	-8450.0	-8700.0	-8625.0	-8662.5	-8825.0	-9100.0	-8100.0	-8850.0	-8225.0	-8662.5	-8500.0
T	-7200.0	-8200.0	-8200.0	-7866.7	-8500.0	-7766.7	-8200.0	-7700.0	-7950.0	-7850.0	-8350.0	-7900.0	-7966.7	-7550.0	-7950.0	-8050.0
V	-8100.0	-8444.4	-8733.3	-8233.3	-8800.0	-8211.1	-8300.0	-8416.7	-8358.3	-8450.0	-8766.7	-7800.0	-8544.4	-7950.0	-8358.3	-8266.7
W	-7400.0	-8000.0	-8300.0	-7700.0	-8150.0	-7750.0	-7850.0	-7850.0	-7850.0	-7775.0	-8225.0	-7550.0	-7950.0	-7475.0	-7850.0	-7925.0
X	-7875.0	-8383.3	-8600.0	-8141.7	-8725.0	-8100.0	-8275.0	-8237.5	-8256.3	-8300.0	-8662.5	-7825.0	-8400.0	-7850.0	-8256.3	-8212.5
Y	-7850.0	-8500.0	-8100.0	-8416.7	-9550.0	-7933.3	-8700.0	-7975.0	-8337.5	-8700.0	-8825.0	-7850.0	-8500.0	-7850.0	-8337.5	-7975.0

Table 3 Unified oligonucleotide dS^0 (cal K^{-1} mol^{-1}) for normal and mixed nucleotides. Horizontal nucleotides for first nucleotide $5'\mathbf{N_1}N_2$ and vertical column for second nucleotide $5'N_1\mathbf{N_2}$.

5'-3'	**A**	**B**	**C**	**D**	**G**	**H**	**K**	**M**	**N**	**R**	**S**	**T**	**V**	**W**	**X**	**Y**
A	-21.3	-21.3	-22.4	-20.9	-21.0	-21.4	-20.7	-21.9	-21.3	-21.2	-21.7	-20.4	-21.6	-20.9	-21.3	-21.4
B	-22.1	-22.4	-22.2	-22.4	-23.3	-22.0	-22.6	-22.1	-22.3	-22.7	-22.7	-21.9	-22.5	-22.0	-22.3	-22.0
C	-22.7	-22.7	-19.9	-23.6	-27.2	-21.2	-24.1	-21.3	-22.7	-25.0	-23.6	-21.0	-23.3	-21.9	-22.7	-20.5
D	-21.6	-22.0	-23.0	-21.5	-21.2	-22.1	-21.4	-22.3	-21.9	-21.4	-22.1	-21.7	-21.9	-21.6	-21.9	-22.3
G	-22.2	-22.2	-24.4	-21.5	-19.9	-23.0	-21.2	-23.3	-22.2	-21.1	-22.2	-22.4	-22.2	-22.3	-22.2	-23.4
H	-21.8	-22.1	-21.5	-22.2	-23.6	-21.5	-22.4	-21.6	-22.0	-22.7	-22.6	-21.2	-22.3	-21.5	-22.0	-21.4
K	-21.8	-22.3	-23.3	-21.8	-21.3	-22.5	-21.8	-22.5	-22.2	-21.5	-22.3	-22.3	-22.1	-22.0	-22.2	-22.8
M	-22.0	-22.0	-21.2	-22.3	-24.1	-21.3	-22.4	-21.6	-22.0	-23.1	-22.6	-20.7	-22.4	-21.4	-22.0	-20.9
N	-21.9	-22.1	-22.2	-22.0	-22.7	-21.9	-22.1	-22.1	-22.1	-22.3	-22.5	-21.5	-22.3	-21.7	-22.1	-21.9
R	-21.8	-21.8	-23.4	-21.2	-20.5	-22.2	-20.9	-22.6	-21.8	-21.1	-21.9	-21.4	-21.9	-21.6	-21.8	-22.4
S	-22.5	-22.5	-22.2	-22.6	-23.6	-22.1	-22.6	-22.3	-22.5	-23.0	-22.9	-21.7	-22.7	-22.1	-22.5	-21.9
T	-21.3	-22.4	-22.2	-22.1	-22.7	-21.9	-22.5	-21.8	-22.1	-22.0	-22.5	-22.2	-22.1	-21.8	-22.1	-22.2
V	-22.1	-22.1	-22.2	-22.0	-22.7	-21.9	-22.0	-22.2	-22.1	-22.4	-22.5	-21.3	-22.3	-21.7	-22.1	-21.8
W	-21.3	-21.8	-22.3	-21.5	-21.9	-21.6	-21.6	-21.8	-21.7	-21.6	-22.1	-21.3	-21.8	-21.3	-21.7	-21.8
X	-21.9	-22.1	-22.2	-22.0	-22.7	-21.9	-22.1	-22.1	-22.1	-22.3	-22.5	-21.5	-22.3	-21.7	-22.1	-21.9
Y	-22.0	-22.5	-21.1	-22.9	-25.0	-21.6	-23.3	-21.5	-22.4	-23.5	-23.0	-21.6	-22.7	-21.8	-22.4	-21.3

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 default 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 nucleotide 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 sequence” 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.

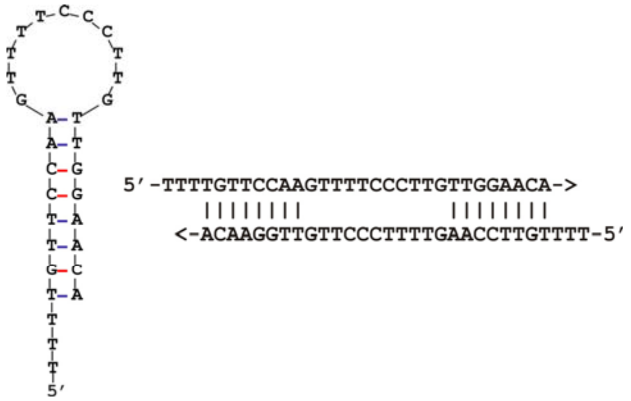
5′
5′-TTTTGTTCCAAGTTTTCCCTTGTTGGAACA->
<-ACAAGGTTGTTCCCTTTTGAACCTTGTTTT-5′

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).

PCR amplification protocol

High quality primers help to perform PCRs even under suboptimal 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 in 25 µl reactions containing 25-50 ng DNA, 1x PCR buffer (with 2 mM $MgCl_2$ or $MgSO_4$), 0.1-1 µM of primer (for primer combinations, we recommend a maximum 1 µM total primer concentration), 0.2 mM dNTPs, 0.5-1 U *Taq* DNA polymerase and optimal additional 0.04 U *Pfu* DNA Polymerase (for long product amplification). A PCR program that suits amplification of both short and long (100-3000 bp) amplicons consists of : 1 cycle at 95°C (3 min), 28-30 cycles of 95°C (10 sec), 60°C (1-60 sec), 68°C (1-60 sec), and a final extension step at 68°C (5 min). We routinely perform PCR amplification in a PTC-100 Programmable Thermal Controller (MJ Research Inc., Bio-Rad Laboratories, USA) or a MasterCycler Gradient (Eppendorf AG, Germany) using 0.2 ml tubes or 96-well plates. General PCR and qPCR set-up can be found at the following site:

http://primerdigital.com/Tools/ReactionMixture.html.

The secondary (non-specific) binding test (alternative amplification)

The specificity of the oligonucleotides is one of the most important factors for good PCR; optimal primers should hybridize only to the target sequence. Particularly when genomic DNA is used as the template, it is possible that primers may bind to inverted repeat regions of the genome. Additional amplification problems can arise due to non-specific primer design for repeated sequences (retrotransposons, SINE, LINE or tandem repeats). Alternative product amplification during PCR also occurs when primer sequences are complementary to an inverted repeat and will produce multiple bands. This is unlikely to occur when primers have been designed using specific DNA sequences (unique PCR). However, the generation of inverted repeat sequences are exploited in two common generic DNA fingerprinting methods – RAPD and AP-PCR (Welsh and McClelland 1990; Williams *et al.* 1990). Since only one primer is used in these PCR reactions, the ends of the products must be reverse complements and thus can form stem-loops.

In plants, the techniques of inter-repeat amplification polymorphisms (IRAP), retrotransposon microsatellite amplification polymorphisms (REMAP) or inter-MITE amplification (Bureau and Wessler 1994; Kalendar *et al.* 1999; Chang *et al.* 2001; Kalendar and Schulman 2006) have exploited the highly abundant dispersed repeats such as the LTRs of retrotransposons and SINE-like sequences. The association of these sequences with each other makes it possible to amplify a series of bands (DNA fingerprints) using primers homologous to these high copy number repeats. The markers generated are very informative as genetic markers.

These methods demonstrate that for repeated sequences there is a probability that single primers can produce PCR products from genomic templates. Even today, we observe that the most useful and popular web 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), such as *Alu* repeats, are dispersed throughout the genome. Primer sequences complementary to any of these repeats may produce many non-specific bands in single-primer amplification. These can be used as markers for detecting *Alu*-repeat polymorphisms (Nelson *et al.* 1989; Sinnet *et al.* 1990).

High copy number repeats will compromise the performance of PCRs so, unless the experiments are planned specifically to exploit them, they are best avoided. A homology search of the primer sequence, for example using 'blastn' against all sequences in a public genome database, e.g, the National Center for Biotechnology Information (NCBI), will determine whether the primer is likely to interact with dispersed repeats. However, in our opinion, the time and effort are not justified: almost all problems of secondary (non-specific) primer binding sites can be simply solved using DNA within BAC clones (about 100,000 bases) rather than using genomic DNA. Another possible way to solve this problem is to create a small 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. Primers detected 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′ base from which 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 secondary non-specific primer binding test is based on a quick, non-gapped local alignment screening between the reference and input sequences.

The alignment for finding 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 $MgCl_2$) 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 screening library dataset (from loading the dataset 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 PCR primer pairs for the given target sequences. The speed of calculation depends on the numbers of target sequences and primer pairs for each of them.

An alternative way to design compatible multiplex PCR primer pairs is to use predesigned 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 or design using common options: the individual setting has higher priority for PCR primer or probe design than the general settings. The results include primers for individual sequences, and compatible primers for PCR, the product sizes, annealing temperatures and the preferred compatible primer with all of the above information.

As clear differentiation and analyses of the products are dependent on using compatible primer pairs in the single reactions, the program recovers all potential variants of primer combinations for analyses of the DNA regions and demonstrates, in tabular form, their compatibility (primer-dimers cross-hybridization, product size overlaps and similar alternative primer pairs based on Tm). The user may choose those alternative primer pair combinations that satisfy primer pair compatibility and length of PCR products. The multiplex PCR algorithm is based on the fast not-recursion method, with the software performing checks on product size compatibility (optimal) and cross-dimer interaction for all primers. The information about primer cross dimers is recorded in the memory to avoid repeating this task in subsequent analyses.

Ideally, all primer pairs in a single reaction should have near equal Ta's. For most multiplex PCRs, there is usually a small variation (up 5°C) between the optimal Ta's of all primer pairs and PCR products. The annealing temperature must be optimal in order to maximise the likelihood of amplifying the target genomic sequences while minimizing the risk of non-specific amplification. To amplify the target genomic sequence effectively, the primer quality and properties should be as high as possible.

We can confirm that FastPCR provides reliable primer pairs for simplex and multiplex PCR experiments to amplify a target gene. Using the program, researchers can select pre-designed primer pairs from a target for their desired types of PCR reactions by changing the filtering conditions as mentioned above. For example, a conventional multiplex PCR requires differently sized amplicons for a set of target genes (at least a 10 bp difference), so the value for the minimum size difference between PCR products can be selected. In addition to the requirement of not having the same sized amplicons, multiplex PCR has other considerations such as minimizing primer dimers and secondary products that become more difficult with increasing numbers of primers involved in a reaction. To avoid the problem of non-specific amplification, FastPCR allows the selection of primer pairs that give the most likelihood of producing only the amplicons of the target sequences by choosing the sequences which avoid repeats or other motifs.

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 for the whole primer sequence. The 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 such as disease resistance analogues (NBS-profiling), transposable elements or other repeat elements (sequence-specific amplification polymorphism, S-SAP) 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 visually 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 options. 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 dye and a quencher moiety: the proximity of the quencher extinguishes the fluorescence of the reporter. When the probe is bound to its complementary single-stranded target DNA it is in the way of the polymerase that is extending from the PCR primer and is degraded by the 5′-3′ exonuclease activity of the enzyme. The reporter is dequenched by their physical separation, leading to increase in the fluorescence signal that is measured as an indication of product accumulation.

There are other fluorescent RT-PCR systems and each method may have its own special properties. For example, the self-reporting primers as described by Fiadaca *et al.* (2001) use three oligonucleotides - a quencher-labeled peptide (or DNA) 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 of polymerase chain reactions 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. 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.

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 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 permutetions 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 fulfils the set similarity parameters. The extension is done by pairwise character comparisons. 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 kmers 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 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.

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

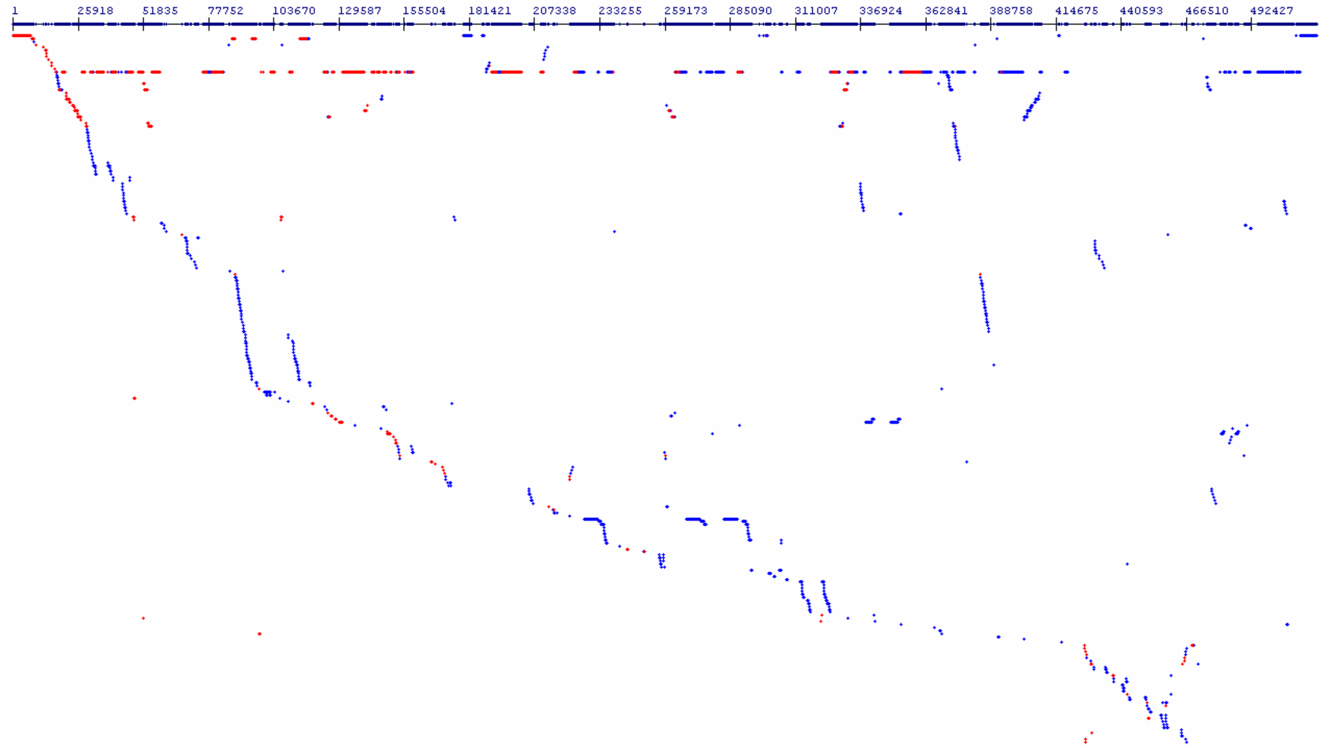

Fig. 3 Results of repeat search using FastPCR. The output shows the result of repeat searching in BAC 438,828 bp *Triticum monococcum* genomic sequence (AY485644; Yan *et al*. 2002) and presented in a graphical display demonstrating that 65% of the genomic sequence is covered with repeated sequences). 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.

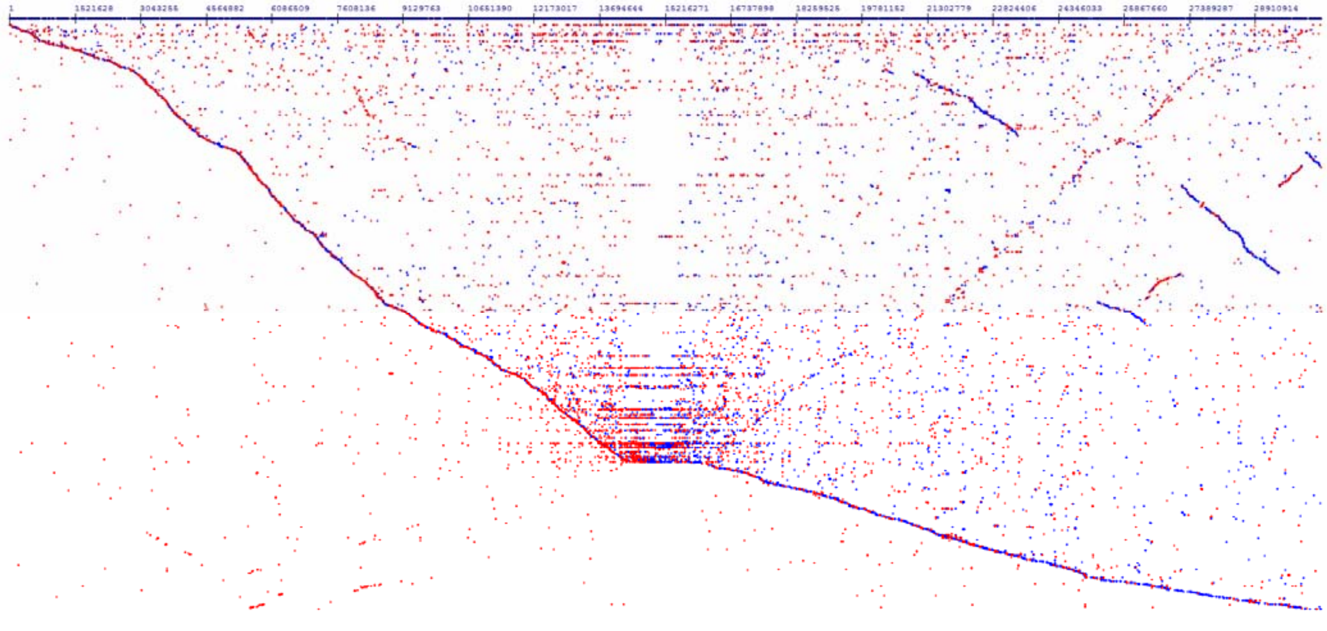

Fig. 4 Show that 26.5% from 30.4M bp sequence of *Arabidopsis thaliana* chromosome 1 (available online: ftp://ftp.arabidopsis.org/home/tair/Sequences/) covering by repeat sequences. The centromere of chromosome 1 is easy to detect in the middle of picture by the special structure of the clusters of centromeric repeats.

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′ 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 polypurine tract (PPT). Some groups of LTR retrotransposons, such as TRIMs and LARDs, have very short internal regions incapable of encoding any proteins; these retrotransposons 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 (Rho *et al*. 2007), and ‘LTRharvest’ (Ellinghaus *et al*. 2008). In all these packages, the idea for discovering complete LTR retrotransposons is to search for direct repeats as representatives of LTRs. From these sequences, the programs extend in both directions until the homology ends. The ends of LTRs are bounded by the sequences TG…CA and flanked by short direct repeats, 4-6 bases. These repeats of the target sequences, called target site duplications (TSDs), are produced during the transposition process by repair of a staggered cut. Between the LTRs are the internal domains of the retrotransposons including the PBS, coding sequences, and PPT.

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 retrotransposons (Marquet *et al*. 1995; Mak and Kleiman 1997; Kelly *et al*. 2003; LeGrice 2003). The PBS of LTR retrotransposons and all retroviruses are complementary to one of a limited set of tRNAs: tRNA^{iMet}, tRNA^{Lys}, tRNA^{Pro}, tRNA^{Trp}, tRNA^{Asn}, tRNA^{Ser}, tRNA^{Arg}, tRNA^{Phe}, tRNA^{Leu}, or tRNA^{Gln}. The tRNA choice appears to depend on the group and one tRNA may predominate: lentiviruses, including HIV-1, use tRNA^{Lys} (Mak and Kleiman 1997), whereas tRNA^{iMet} is commonly, though not exclusively, used by plant retrotransposons. A handful of exceptional LTR retrotransposons, including the *Tf1/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; Kalendar *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 the PBS-like domains. We were able to identify *in silico* 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 k-mer seeds is changeable (default = 40 bases with 70% similarity). The algorithm does not remove low-complexity sequences, because many LTRs contain them.

In the second phase, the algorithm searches for probable PBS sites (12 bases) allowing 2 mismatches with the complementary 3′ tRNA sequences. PBS sequences were designed using sequences data from public and local LTR retrotransposon databases (Kalendar *et al*. 2009). The PBS site helps to locate 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 haphazard 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 of a putative LTR retrotransposon. The solution is partly manual analysis of each individual LTR retrotransposon. By collecting all full-size and partial sequences of 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; Kalendar *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* RetrOryza 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; Kalendar *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 regions 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 provides a means for the design 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 comprehensive 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

Abajian C (1994) SPUTNIK. Available online: http://espressosoftware.com/pages/sputnik.jsp

Allawi HT, SantaLucia J (1997) Thermodynamics and NMR of internal G-T mismatches in DNA. *Biochemistry* **36**, 10581-10594

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* **12**, 1269-1276

Bekaert M, Teeling EC (2008) UniPrime: a workflow-based platform for improved universal primer design. *Nucleic Acids Research* **36**, e56

Benjak A, Forneck A, Casacuberta JM (2008) Genome-wide analysis of the "cut-and-paste" transposons of grapevine. *PLoS ONE* **3**, e3107

Bommarito S, Peyret N, SantaLucia JJ (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Research* **28**, 929-1934

Boutros PC, Okey AB (2004) PUNS: transcriptomic and genomic in silico PCR for enhanced primer design. *Bioinformatics* **20**, 2399-2400

Bureau TE, Wessler SR (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**, 907-916

Campagna D, Romualdi C, Vitulo N, Del Favero M, Lexa M, Cannata N, Valle G (2005) RAP: a new computer program for *de novo* identification of repeated sequences in whole genomes. *Bioinformatics* **21**, 582-588

Cao Y, Wang L, Xu K, Kou C, Zhang Y, Wei G, He J, Wang Y, Zhao L (2005) Information theory-based algorithm for *in silico* prediction of PCR products with whole genomic sequences as templates. *BMC Bioinformatics* **6**, 190

Chang RY, L. O'Donoughue L, Bureau TE (2001) Inter-MITE polymorphisms (IMP): a high throughput transposon-based genome mapping and fingerprinting approach. *Theoretical and Applied Genetics* **102**, 773-781

da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A (2008) SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* **2008**, 412696

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* **21 (Suppl. 1)**, i152-8

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18

Fiandaca MJ, Hyldig-Nielsen JJ, Gildea BD, Coull JM (2001) Self-reporting PNA/DNA primers for PCR analysis. *Genome Research* **11**, 609-613

Flavell AJ, Dunbar E, Raymond Anderson, Pearce SR, Hartley R, Kumar A (1992) Ty1–copia group retrotransposons are ubiquitous and heterogeneous in higher plants. *Nucleic Acids Research* **20**, 3639-3644

Fredslund J, Lange M (2007) Primique: automatic design of specific PCR primers for each sequence in a family. *BMC Bioinformatics* **8**, 369

Gabrielian A, Bolshoy A (1999) Sequence complexity and DNA curvature. *Computer and Chemistry* **23**, 263-274

Gadberry MD, Malcomber ST, Doust AN, Kellogg EA (2005) Primaclade - a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics* **21**, 1263-1264(2)

Gilson MK, Given JA, Bush BL, McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal* **72**, 1047-1069

Hizi A (2008) The reverse transcriptase of the Tf1 retrotransposon has a specific novel activity for generating the RNA self-primer that is functional in cDNA synthesis. *Journal of Virology* **82**, 10906-10

Jewell E, Robinson A, Savage D, Erwin T, Love CG, Lim GA, Li X, Batley J, Spangenberg GC, Edwards D (2006) SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery. *Nucleic Acids Research* **1**, W656-659

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462-467

Kalendar R, Antonius K, Smýkal P, Schulman AH (2009) iPBS amplification, a universal method for isolating retrotransposons and displaying insertional polymorphisms. *Plant Physiology*, submitted.

Kalendar R, Grob T, Regina MT, Suoniemi A, Schulman AH (1999) IRAP and REMAP: Two new retrotransposon-based DNA fingerprinting techniques. *Theoretical and Applied Genetics* **98**, 704-711

Kalendar R, Schulman AH (2006) IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nature Protocols* **1**, 2478-2484

Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O, Schulman AH (2008) Cassandra retrotransposons carry independently transcribed 5S RNA. *Proceedings of the National Academy of Sciences USA* **105**, 5833-5838

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* **166**, 1437-1450.

Kalyanaraman A, Aluru S (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. *Journal of Bioinformatics and Computational Biology* **4**, 197-216

Kelly NJ, Palmer MT, Morrow CD (2003) Selection of retroviral reverse transcription primer is coordinated with tRNA biogenesis. *Journal of Virology* **77**, 8695-8701

Kurtz S (2000) The VMATCH large scale sequence analysis software. Available online: http://www.vmatch.de

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**, 4633-4642

Lefebvre A, Lecroq T, Dauchel H, Alexandre J (2003) FORRepeats: detects repeats on entire chromosomes and between genomes. *Bioinformatics* **19**, 319-326

LeGrice SFJ (2003) "In the Beginning": Initiation of Minus Strand DNA Synthesis in Retroviruses and LTR-Containing Retrotransposons. *Biochemistry* **42**, 14349-14355

Lexa M, Horak J, Brzobohaty B (2001) Virtual PCR. *Bioinformatics* **17**, 192-193

Lexa M, Valle G (2003) PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics* **19**, 2486-2488

Mak J, Kleiman L (1997) Primer tRNAs for reverse transcription *Journal of Virology* **71**, 8087-8095

Marquet R, Isel C, Ehresmann C, Ehresmann B (1995) tRNAs as primer of reverse transcriptases. *Biochimie* **77**, 113-124

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362-367

Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211-218

Nazarenko I, Lowe B, Darfler M, Ikonomi P, Schuster D, Rashtchian A (2002) Multiplex quantitative PCR using self-quenched primers labeled with a single fluorophore. *Nucleic Acids Research* **30**, e37

Nelson DL, Ledbetter SA, Corbo L, Victoria MF, Ramirez-Solis R, Webster TD, Ledbetter DH, Caskey CT (1989) Alu polymerase chain reaction: a method for rapid isolation of human specific DNA sequences from complex DNA sources. *Proceedings of the National Academy of Sciences USA* **86**, 6686-6690

Nishigaki K, Saito A, Takashi H, Naimuddin M (2000) Whole genome sequence-enabled prediction of sequences performed for random PCR products of Escherichia coli. *Nucleic Acids Research* **28**, 1879-1884

Novère Le (2001) MELTING, a free tool to compute the melting temperature of nucleic acid duplex. *Bioinformatics* **17**, 1226-1227

Orlov YL, Potapov VN (2004) Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Research* **32**, W628-W633

Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, Pedersen KF, Lin Y, Garretson J, McEntaggart NO, Sailor CA, Dawson RB, Peek AS (2008) IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Research* **1**, **(Web Server issue)**, W163-169

Peyret N, Seneviratne PA, Allawi HT, SantaLucia J Jr. (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

Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics* **8**, 90

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (Eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, Humana Press, Totowa, NJ, USA, pp 365-386

Rubin E, Levy AA (1996) A mathematical model and a computerized simulation of PCR using complex templates. *Nucleic Acids Research* **24**, 3538-3545

Rychlik W, Spencer WJ, Rhoads RE (1990) Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucleic Acids Research* **18**, 6409-6412

SantaLucia JJ (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbour thermodynamics. *Proceedings of the National Academy of Sciences USA* **95**, 1460-1465

Sinnet D, Deragon J-M, Simard LR, Labuda D (1990) Alumorphs–human DNA polymorphisms detected by polymerase chain reaction using Alu-specific primers. *Genomics* **7**, 331-334

Sugimoto N, Nakano S, Yoneyama M, Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Research* **24**, 4501-4505

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* **18**, 7213-7218

Williams JGK, Kubelik AR, Livak KL, Rafalscki 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

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35 (Web Server issue)**, W265-268

Yan L, Echenique V, Busso C, SanMiguel P, Ramakrishna W, Bennetzen JL, Harrington S, Dubcovsky J (2002) Cereal genes similar to *Snf2* define a new subfamily that includes human and mouse genes. *Molecular Genetics and Genomics* **268**, 488-99

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

Yoshimura M, Nakamura S, Ogawa H (2005) TaqMan Real-Time PCR quantification conventional and modified methods. *Methods in Molecular Medicine* **108**, 189-198