DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
141

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
141

METHODS AND SOFTWARE FOR PREDICTING PCR FAILURE RATE IN LARGE GENOMES

REIDAR ANDRESON

TARTU UNIVERSITY
PRESS

Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of *Doctor philosophiae* (in bioinformatics) on 14.04.2008 by the Council of the Institute of Molecular and Cell Biology, University of Tartu.

Supervisor: Prof. Maido Remm, PhD (University of Tartu, Estonia)

Opponent: Prof. Dong Xu, PhD (University of Missouri-Columbia, USA)

Commencement: Room No 217, 23 Riia Str., Tartu, on June 11th, at 14.00.

The publication of this dissertation is granted by the University of Tartu.

The doctoral studies and the publication of the current thesis were supported by the Graduate School of Biomedicine and Biotechnology created under the auspices of European Union Social Fund structural funds measure 1.1. *Educational System Supporting the Flexibility and Employability of the Labor force and Providing Opportunities of Lifelong Learning for All.*

ISSN 1024–6479
ISBN 978–9949–11–855–7 (trükis)
ISBN 978–9949–11–856–4 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr 193

TABLE OF CONTENTS

2

LIST OF ORIGINAL PUBLICATIONS

The current dissertation is based on the following publications referred to in the text by their Roman numerals:

I. Kaplinski L, **Andreson R**, Puurand T, Remm M (2005). MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics* 21(8): 1701–2.

II. **Andreson R**, Reppo E, Kaplinski L, Remm M (2006). GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics* 7:172.

III. **Andreson R**, Puurand T, Remm M (2006). SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. *Nucleic Acids Research* 34:W651–5.

IV. **Andreson R**, Möls T, Remm M (2008). Predicting failure rate of PCR in large genomes. *Nucleic Acids Research* (accepted)

My contribution to the articles referred in the current thesis is as follows:

Ref. I created and performed tests with auxiliary program *gt4multiplx* and also participated in creation of the web client for the main program;
Ref. II conducted this study, carried out different tests on various methods, validated the package and was responsible for drafting the manuscript;
Ref. III conducted this study, created and performed tests with the application and was responsible for drafting the manuscript;
Ref. IV conducted this study, created primers for the experiments, analyzed the data and was responsible for drafting the manuscript.

LIST OF ABBREVIATIONS

CEPH	Centre d'Etude du Polymorphisme Humain
CPU	central processing unit (in computers)
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide-triphosphate
e-PCR	electronic (*in silico*) PCR
IUPAC	International Union for Pure and Applied Chemistry
n-mer	short substring with the length of n
LINE	long interspersed nuclear element
LTR	long terminal repeat
PCR	polymerase chain reaction
RAM	random access memory (in computers)
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
T_a	primer annealing temperature
T_m	primer melting temperature

INTRODUCTION

Modern genomic technologies allow studying thousands of genomic regions from each DNA sample. Many of these technologies rely on methodology called polymerase chain reaction (PCR) that allows amplification of specific DNA sequences (gene detection for example). The genome-wide genotyping of single nucleotide polymorphisms (SNP), microarray experiments for gene expression, re-sequencing methods – all these depend directly on the efficiency of the PCR reaction. The high-throughput assays require designing simultaneously thousands of PCR primers for the experiments. Therefore, careful estimation of the PCR primer properties is crucial for the success of primer design. Many studies in the past have been focused on optimizing the reagents of the PCR reaction such as concentration of reaction buffer components and PCR protocols. On the other side there is a primer design process. The basic oligonucleotides properties and their optimal combinations are well studied by many scientific groups in order to maximize the amplification efficiency. However, the in-depth examination of the repeats and the uniqueness of PCR primers in large genomes are still under discussion.

The first part of the present thesis gives a brief overview of the PCR method, both biochemical and sequence-based factors influencing the PCR reaction and studies to measure the effects of these factors. The second major topic of the literature review concentrates on the eukaryotic repeats, their classification and methods to detect them. Third part gives an overview of the current electronic PCR (e-PCR) methods that are available today.

The research part of this dissertation entails the following topics: (i) creation of the fast and efficient repeat-masking methodology designed for PCR applications, (ii) creation of the fast and brute-force method to predict PCR products for already designed PCR primers and (iii) discovery of the important factors that affect the PCR failure rate and create statistical models to predict the failure rate of PCR reaction.

1. REVIEW OF LITERATURE

1.1. Polymerase chain reaction

1.1.1. The essence of PCR

The Polymerase Chain Reaction (PCR) technique, conceived by Kary B. Mullis, allowed scientists to make millions of copies of a slight amount of DNA (Saiki *et al.*, 1985, Mullis *et al.*, 1986). This technique, in vitro DNA amplification procedure, has been optimized, improved and perfected in the following years (Saiki *et al.*, 1986, Scharf *et al.*, 1986, Mullis and Faloona, 1987, Saiki *et al.*, 1988, Lawyer *et al.*, 1989, Olson *et al.*, 1989, Erlich *et al.*, 1991). Furthermore, the PCR has revolutionized many aspects of the research ever since and *Science* has nominated in 1989 the DNA polymerase to be the "Molecule of the Year" based on the accomplishments of PCR method (Guyer and Koshland, 1989), for which Kary Mullis was awarded the 1993 year's Nobel prize in Chemistry.

The PCR reaction itself is based on the cyclic synthesis of both DNA chains. A standard PCR amplification involves three following steps: heat denaturation of double-stranded DNA, annealing of the two primers (short oligonucleotides) to their complementary sequences and extension of the annealed primers with thermostable DNA polymerase. An ideal ordinary PCR result is one specific PCR product that is generated in high yield, with minimal cycles containing the fewest number of polymerase-induced errors. The amount of amplified PCR product is doubled in each successive cycle causing the exponential accumulation of given specific fragment (Saiki *et al.*, 1988).

Nowadays there are more advanced PCR technologies, such as Real-Time PCR (Higuchi *et al.*, 1993, Heid *et al.*, 1996), that are commonly utilized in current research projects. In classical PCR the same amount of product is produced independently of the initial amount of DNA template molecules. In real-time PCR however, the number of amplification cycles required to obtain a particular amount of DNA molecules is registered by monitoring the fluorescence of dyes or probes introduced into the reaction (Kubista *et al.*, 2006). This data can be analyzed by computer to calculate the amount of product formed during each reaction cycle. Nevertheless, classical PCR technique is still widely used on many fields due to its efficiency, robustness and fidelity (Vollenhofer *et al.*, 1999, Kurg *et al.*, 2000, Nugent and Saville, 2004, Budowle *et al.*, 2005, Yancy *et al.*, 2005).

3

1.1.2. The estimation of PCR success

Although the PCR methodology is evolved and protocols are optimized by decades now, the behavior of the reaction is not completely predictable for each new primer-template combination. The non-successful results of a classical PCR include non-targeted products, smear bands or no bands at all. The alternative products are mostly caused by non-unique PCR primers that amplify additional regions from template DNA. The reasons for other non-successful results may be either sequence-based or experimental errors. A closer look to these factors is given in the next chapter.

In order to achieve a high PCR success rates, primers need to be selected carefully. In the beginning of the PCR "era", researchers were amplifying sequences from less complex organisms such as microbes and viruses. Today, with the advancement of genome sequencing project, the genomic DNAs of several higher organisms (like mammals, plants) are available and therefore the specificity of PCR primers requires much closer attention. Even though the cost of single PCR reaction is comparatively low, it is becoming an issue in high-throughput methods for genomic applications.

The prediction of the success of PCR has been studied previously by many groups. Rubin and Levy published a study, where they investigated the relative effects of various parameters on the amplification of non-targeted PCR products (Rubin and Levy, 1996). The most significant factor affecting the PCR specificity is the mismatch tolerance during primer annealing to the template, followed by primer length, template size and product size limits. Beasley with her colleagues have analyzed a thousands of primer pairs and examined the primer characteristics that can cause a false priming or failure to amplify template DNA (Beasley *et al.*, 1999). They have found that the primer length, primer GC content and GC content of the 3' half of the primers were strongly associated with the success rate of PCR. Yuryev along with his workgroup developed statistical scores to evaluate various parameters for predicting the success of primer extension reaction that includes many factors related to PCR primers and products (Yuryev *et al.*, 2002). The statistical prediction (single-plex) model included following PCR-related factors: primer GC content, the number of ambiguous bases and repeats in PCR product, the product structure around PCR primer annealing sites and the nucleotide combinations in last 3 bases at the 3' end of the PCR primers in addition to two product bases next to primer annealing sites. The PCR success can be predicted by the regionalized GC content within the template DNA (Benita *et al.*, 2003). Benita with the co-authors has published a detailed analysis of the template DNA using a sliding window of 21 nucleotides to calculate GC nucleotides in each window. Region was considered significant when it contains >61% GC nucleotides for at least ten consecutive windows. These threshold values gave more precise discrimination between 'good' and 'failed' experiments than any other parameters they have used. A critical examination of oligonucleotides properties has been

published lately (Chavali *et al.*, 2005). The authors propose that the efficiency and accuracy of the PCR are determined by correct calculation of the primer melting temperature (T_m) and secondary structures. They compare several freely available programs and provide suggestions to use different tools depending on the template GC content. The factors affecting cross-species primers and their success in PCR has been studied by Housley and her colleagues (Housley *et al.*, 2006). They have identified three factors with significant impact on the efficiency of PCR: the number of index-species mismatches, GC content of the template and the degree of relatedness between two organisms.

1.1.3. The factors influencing PCR

There are many factors that affect the success of the PCR and can be generally divided into two subgroups: experimental or biochemical and sequence-based factors.

1.1.3.1. Experimental and biochemical factors

The optimal selection of PCR reaction components is crucial for running a successful experiment. The correct annealing of two sequences (PCR primer and DNA template) to each other does, however, depend on the physical and chemical solution conditions under which the reaction takes place. The recommended PCR buffer reagents and their concentrations have been published previously (Innis and Gelfand, 1990). Although the modern formulations may differ considerably, they are generally comparable. Magnesium ion concentration influences many things in the reaction: primer annealing, T_m of template (strongly influences ΔS), product and primer-template associations (high magnesium will enhance the stability of mismatched primers), the enzyme activity and fidelity (important cofactor for Taq DNA polymerase). A titration should be performed with varying [Mg^{++}] with all new template-primer combinations as the results can differ markedly even under the same conditions of concentrations and cycling times/temperatures. Primer and deoxynucleotide triphosphates (dNTPs) concentrations should also not be too high; 0.2µM should be more than sufficient for homologous primers and <50µM for each dNTP (Innis and Gelfand, 1990, Beasley *et al.*, 1999).

The PCR cycle includes 3 steps: denaturation of double-stranded DNA, annealing of the primer sequences to single-stranded DNA template and synthesis of a new complementary strand on the template. A typical DNA template denaturing temperature is set between 93º and 96ºC for every cycle of an amplicon. In the denaturation step the Taq DNA polymerase is inactivated and eventually will lose its activity. After 10 cycles the amplified product acts as a template and therefore it is unnecessary to use same temperature during later cycles. For short amplicons it is proposed that the denaturation temperature

should be lowered to 87º–90ºC after five to ten initial cycles (Yap and McGee, 1991). The increase in denaturation temperature and decrease in time may also work (96^{o}C for 15 sec) (Innis and Gelfand, 1990).

The annealing temperature (T_a) of the PCR primer is related to the T_m and one should aim the T_a about 5^{o}C below the lowest T_m of the pair of primers. Thus, the correct T_m prediction is needed in order to get the precise T_a for the given primer sequence (SantaLucia, 1998, von Ahsen *et al.*, 2001). Too low T_a increases a chance that one or both primers will anneal to sequences other than the true target**,** as internal single-base mismatches or partial annealing of the primers may be tolerated. Too high T_a, on the other hand, can cause the deficiency of the synthesized product, as the likelihood of primer annealing is reduced.

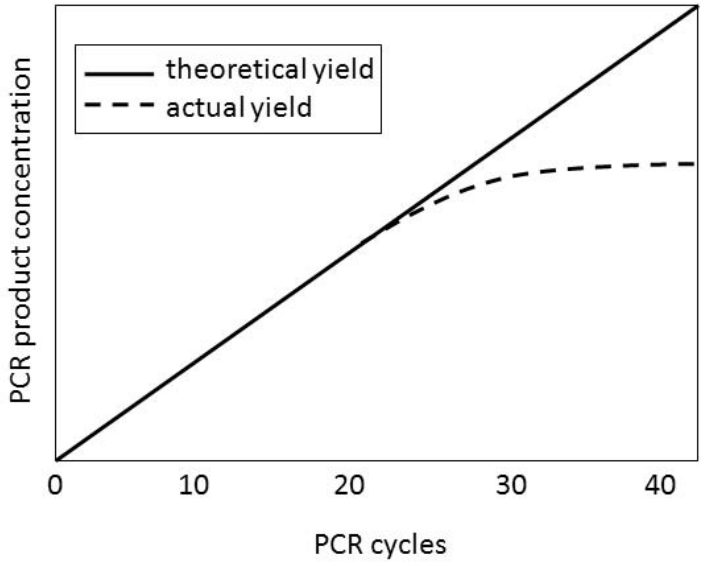


Figure 1. "Plateau effect" in PCR amplification. The attenuation in the exponential rate of PCR product accumulation happens in the late stages of a PCR due to degradation of reactants or reactant depletion.

The total number of cycles depends on the concentration of target molecules: from 40–45 cycles to amplify 50 target molecules to 25–30 to amplify 3×105 molecules to the same concentration. On both cases the exponential growth of the product will diminish (Figure 1) at some stage caused by degradation of reactants (dNTPs, enzyme) with short products, reactant depletion (primers, dNTPs) with long products or competition for reactants by non-specific products (Innis and Gelfand, 1990).

1.1.3.2. Sequence-based factors

The sequence specific factors affecting the PCR success rate can be divided additionally into following groups: PCR primer and product properties (e.g. length, GC content), PCR primer secondary binding sites, primer homology on the target DNA and the number of alternative PCR products. In case of primer design the calculation of the properties in first two groups is an order of magnitude faster than providing the uniqueness of primers and products with whole genomes (data not shown). Furthermore, it requires more computing power, space and sophisticated algorithmic approaches to accomplish latter tasks. Therefore, the special chapters are devoted for masking repeats and counting PCR products afterwards.

PCR primer properties

The first PCR primer property, PCR primer length, is dependent on the base composition and the melting temperature set by the researcher. A prime consideration is that the primers should be complex enough so that the likelihood of primer annealing to sequences other than the chosen target is very low and as short as possible to lower the cost of the primer synthesis. The primer sequence containing sixteen nucleotides will statistically be present only once in every 4^{16} bases (>4 billion) and should be theoretically unique in human genome. Furthermore, it is shown that primers with lengths between 21–26 nucleotides give higher success rates than shorter (18–20 nt) (Beasley *et al.*, 1999).

The extreme GC content (>80 or <20%) of the full primer sequence and in the 3' half of the primer are known to increase the probability of self-complementarities and secondary binding sites. Beasley *et al.* and Haas *et al.* recommend to use primers with GC content close to 50% (Haas *et al.*, 1998, Beasley *et al.*, 1999).

The successful elongation of a primer depends also on the stability at its 3' end (Onodera and Melcher, 2004, Miura *et al.*, 2005). It is shown that primers with a G or C in the last base at 3' end are more likely to succeed in PCR (Li *et al.*, 1997, Onodera and Melcher, 2004). More specifically, the last base should not be a Thymine (T) because of its ability to form non-Watson-Crick base pairs (mismatch tolerance) and increase the probability of secondary binding sites (Kwok *et al.*, 1990). As DNA polymerases are known to form a duplex between not identical primer and template sequences, there are programs available that evaluate the duplex energy of 3' half of the primer candidates (Haas *et al.*, 1998, Rozen and Skaletsky, 2000, Chen *et al.*, 2003, Miura *et al.*, 2005).

The short simple repeats in primer sequence can cause a higher probability of having stable secondary binding sites in genomic DNA (Haas *et al.*, 1998). Figure 2 shows some examples of primers containing simple repeats; a more detailed overview of the repeats and masking methods is given in the next chapter.

```
>primer1
CTGCTTaaaaaaaaaGTaaaaaaaa

>primer2
AATCAAGActctctctctctctGAA

>primer3
ctgctgGTTCAGGCAActgctgctgC
```

Figure 2. Examples of short repeats in primer sequences. Lower-case letters mark the simple repeats.

The avoidance of the formation of primer-dimer artefacts (Figure 3A and 3B) and stable self-complementary hairpin loops (Figure 3C) that compete with the correct primer-template target binding are important to increase the specificity of PCR primers (Chavali *et al.*, 2005).

A. primer-dimer

```
dG: -5.4 kcal/mol
    5'-AGCTTGGGAGCTAGCCAAAC-3'
             ||||
3'-CAATAGTAGCCCTACCTTG-5'
```

B. primer-dimer 3' ends

```
dG: -5.8 kcal/mol
5'-ATGGCTTCTAGAGGACCAATG-3'
                   |||||
                3'-GTTACCTTTAACGCAGGTATC-5'
```

C. hairpin

```
dG: -1.64 kcal/mol
                T
               / \
        5' AGCT   G
           ||||   |
3' CAAACCGATCGA   G
               \ /
                G
```

Figure 3. Examples of the secondary structures of PCR primers. The free energies of these secondary structures are calculated using following conditions: DNA at 37°C $[Na^{+}]$ = 0.05 M, $[Mg^{++}]$ = 0.0015 M. The cross dimer (A and B) structures and energies were calculated with MultiPLX (Ref. I) and the hairpin (C) calculated with MFOLD web server (Zuker, 2003).

The important characteristic for the calculation of correct annealing temperature is the melting temperature of the primer as noted above. The choice of a non-optimal temperature can lead to the amplification of false regions. The proper calculation of T_m using latest Nearest-Neighbor thermodynamic formulas (Owczarzy *et al.*, 1997, SantaLucia, 1998, SantaLucia and Hicks, 2004, Panjkovich and Melo, 2005) requires exact concentrations of the molecules from the reaction protocol (von Ahsen *et al.*, 2001, Chavali *et al.*, 2005), thus making the T_m experimentally dependent sequence-based factor.

PCR product properties

The previous studies have shown that the amplicon length is not a critical factor affecting the result of PCR (Beasley *et al.*, 1999, Benita *et al.*, 2003). However, the GC content of the PCR product is more informative (Benita *et al.*, 2003). It has been found that DNA templates with very high or low GC content can be difficult to amplify (Varadaraj and Skinner, 1994). The stable secondary structures of the target DNA to which the primers bind are also important to look after as they obstruct the DNA denaturation and the progress of the polymerase (Fedorova *et al.*, 1992, Dong *et al.*, 2001). Finally, the repetitive elements located on the PCR products are increasing the primer mispriming (Haas *et al.*, 1998).

1.2. Key concepts for masking repeats

1.2.1. Repeats in eukaryotes

Most of the eukaryotic organisms comprise a large fraction of repetitive motifs in their genomic DNA. Current estimates are that 46% of the human and 38% of mouse genomes are occupied by various repeats (Lander *et al.*, 2001, Waterston *et al.*, 2002). These repetitive motifs can be divided roughly into three categories: simple (duplications of simple sets of DNA bases (typically 1–13bp) or minisatellites (14–500bp)), tandem (duplications of more complex 100–200 base sequences) and interspersed repeats (SINEs, LINEs, LTRs and DNA transposons) (Richards and Sutherland, 1994, Prak and Kazazian, 2000, Nagashima *et al.*, 2004).

Although repetitive motifs were once called as a residual “junk DNA”, that opinion is about to change today. It is even argued that repeats probably play an important role of developing the species through genome modifications (Kazazian, 2004). Therefore the role of repeats are noted often as “symbiotic” rather than “parasitic” and the research on this field is an emerging area in evolutionary biology (Zhi *et al.*, 2006).

In addition, accurate identification and classification of repeats is important for developing sequence assembly and genome comparison methods (Edgar and

Myers, 2005), understanding diseases caused by repeats (Deininger and Batzer, 1999) and homology searches and oligo design to avoid the explosion of unnecessary or non-unique results (Kreil *et al.*, 2006).

1.2.2. The masking of repeats

At first we must define "masking" to go further. For example let's say that repeat is one string containing 16 nucleotides and it is presented more than 100 times in genomic DNA in several places. How to mark those places on genomic DNA to make sure that we could recognize them later on? The easiest way to mark them is to replace all nucleotides in length of the repeated string by some other symbol (e.g. "N") than ATGC. When the genomic DNA is scanned through and all repeated strings are replaced, we can say that our sequence is masked.

The main obstacle for masking sequences is the volume of eukaryotic DNA. We cannot simply scan large genomic DNA by brute-force and count or find *de novo* repetitive motifs as it is too time consuming. The other important criterion is the sensitivity of the method. Ideally, all motifs that are defined by given rules as repeats should be found. The sensitivity is a problem of methods, whose algorithms are based on some heuristics. To accept these challenges the specific algorithms are needed.

1.2.3. The methods for finding and masking repeats

There are two separate approaches to locate repeats in biological sequences: using predefined or experimentally verified libraries or trying to find repeats directly from nucleic acid sequence without prior knowledge.

The most widely used program is definitely RepeatMasker (Smit, AFA, Hubley, R and Green, P. http://www.repeatmasker.org/) which uses precompiled representative repeat libraries to run homology search with query sequence. There is also a helper application to speed up RepeatMasker called MaskerAid (Bedell *et al.*, 2000). Instead of using CrossMatch application to find homology between sequence and RepBase repeat library (Jurka, 2000), MaskerAid utilizes WU-BLAST (Gish, W. (1996–2004), http://blast.wustl.edu) for a given task. WU-BLAST is an enhanced version of the original NCBI BLAST (Altschul *et al.*, 1990, Altschul *et al.*, 1997). Replacement of CrossMatch with MaskerAid/WU-BLAST increases the speed of masking more than 30-fold without losing the sensitivity (Bedell *et al.*, 2000). CENSOR (Kohany *et al.*, 2006) is a new tool for identification of both interspersed and tandem repeats using similarity searches with NCBI BLAST or WU-BLAST against RepBase.

DUST (ftp://ncbi.nlm.nih.gov/pub/tatusov/dust/) is a program for filtering low complexity regions from nucleic acid sequences. It catches all repeats of unit length 1 or greater that are repeated at least 4 times. For detecting and masking tandem repeats a program called TandemRepeatFinder is developed by Gary Benson (Benson, 1999). It searches tandem repeat patterns using short substrings (n-mer matches), requires no predefined size and number of the patterns (instead it is using a probabilistic model to calculate them) and determines a single consensus pattern for the smallest repetitive motif in the tandem repeat. The program will find all repeats with period size between 1 and 2000.

All the methods described above are specialized for masking repeats. There exist many alternative applications for finding *de novo* repetitive motifs that do not require predefined repeat libraries. In some decades ago Hugo Martinez and Devereux with his colleagues developed the earliest repeat finding algorithms for molecular biologists (Martinez, 1983, Devereux *et al.*, 1984), but the main problem with those tools was the strict limit on the maximal length of the input sequence they were capable to analyze (Kurtz and Schleiermacher, 1999).

In the present day it is recommended that the programs are able to handle complete genomic DNA when detecting repeats. RepeatMatch (Delcher *et al.*, 1999) performs a maximal unique match decomposition of the two closely related genomes using suffix trees combined with the longest increasing sub-sequence and Smith-Waterman algorithm (Smith and Waterman, 1981). REPuter (Kurtz *et al.*, 2001) can handle effectively large genomes by finding exact repeats in linear space and time using a revised implementation of suffix trees. In the second step the exact matches are used as a seeds and extending them allowing mismatches, insertions and deletions, program guarantees that all repeats will be found according to the user input parameters. RepeatFinder (Volfovsky *et al.*, 2001) is a program designed to find, output detailed classification and statistics of all repeats for partial or complete genomes. The gathering of initial set of exact repeat hits is performed using efficient suffix tree data structure. The second stage is a merging procedure that joins overlapping repeats or repeats with limited distance together. Third step is the classification of newly formed combined repeats and the last (optional) step allows the user to WU-BLAST all similar but non-exact repeats against all others. After the final step repeat classes will be updated and program can build repeat map of the whole genome sequence. The authors of Recon (Bao and Eddy, 2002) propose that the repeat families collected by their application can be used as the basis of creating higher quality libraries such as RepeatMasker library. The algorithm is forming the multiple alignments of repeats with WU-BLAST. FORRepeats (Lefebvre *et al.*, 2003) is using a heuristical approach to minimize the search time and space requirements when using large genomes. Lefebvre and his colleagues are using specific heuristical data structure called *factor oracle* (an automaton) that allows them to perform faster pair-wise alignments of exact repeats. The second step is the extension of exact hits that is

similar to BLAST. On the other hand, Pevzner with his colleagues have shown that neither pair-wise (RepeatMatch, REPuter) nor multiple alignment (RepeatFinder, Recon) methods alone are so successful of classifying repeats as their RepeatGluer (Pevzner *et al.*, 2004). Instead, they are using A-Bruijn graphs to eliminate the "mosaic" nature of the sub-repeats (smaller repeats that are overlapping or part of the bigger repeats). The program creates the matrix of input sequence, constructs the A-Brujin graph and removes bulges, whirls and zigzag patterns from the graph. PILER (Edgar and Myers, 2005) is a program package that is using different search methods for several repeat classes. For finding local and multiple alignments PALS (Edgar and Myers, 2005) and MUSCLE (Edgar, 2004) are used respectively. The output of the PILER is an annotation of the input sequences giving locations of intact, isolated copies of repeated elements and a library containing one consensus sequence for each family. RepeatScout (Price *et al.*, 2005) builds a set of repeat families by using high frequency of short substrings with fixed length as seeds. The next step involves the greedy extension of each seed to a longer consensus sequence. Those sequences are aligned against the genome to locate all repeats.

RAP (Campagna *et al.*, 2005) and WindowMasker (Morgulis *et al.*, 2006) are applications that are based purely on a word-counting algorithms. This is an alternative way to find repeats and rely on the statement that a sequence containing frequent words is very likely a repeat. The former program allows using discontinuous words whereas latter program uses exact words only. The algorithm of both programs is divided into two separate parts: at first it count alls n-mers and then the sequence masking (WindowMasker) or visualization of repeats (RAP) will occur. The input sequence will be scanned two times in both cases. These methods are optimized for short word sizes (16 or less), but WindowMasker counts required word size dynamically unlike RAP, where the user defines it manually. A novel method for finding fragmented repeats is called Greedier (Li *et al.*, 2008). This method is using separate iterations to locate transposons: 1) identifies the local similarities between predefined repeat library and target sequence and 2) computes a fitness value for each match separately to tag repeat motifs. Experiments show that Greedier is twice as effective as WindowMasker or RepeatMasker for finding true positive transposon bases and avoiding false positives.

To conclude the overview of different repeat finding and masking methods, the question how to represent all repeats in genomic sequence is still open. As the Bao and Eddy wrote in their paper (Bao and Eddy, 2002), "The problem of automated repeat sequence classification is inherently messy and ill-defined and does not appear to be amenable to a clean algorithmic attack." Current methods approach differently to the problem, but yet there is no ideal solution or common understanding how to classify and draw borders between repeat candidates. Additionally, there are many programs available for finding and/or masking repeats, but only few of them (DUST, RepeatMasker, WindowMasker)

are practically usable in a large-scale whole genome primer or probe design studies. Therefore the need for fast and specialized tools still exists.

1.3. The electronic PCR

Ten years ago, Gregory Schuler introduced to the scientific community a new term called electronic PCR (e-PCR) (Schuler, 1997). The closer definition for this term is the following: e-PCR is the process of recovering sequence-tagged sites (STSs) in DNA sequences by searching for subsequences that closely match the PCR primers and are in the correct order, orientation and spacing to be consistent with the PCR product size. We are widening the definition of e-PCR by saying that e-PCR is the process of counting all binding sites of PCR primers and possible PCR products they may produce in a given sequence within a certain distance.

1.3.1. The relevant e-PCR methods

Many current probe and PCR primer design applications use various mechanisms to exclude non-unique oligo candidates from the regions of interest. Some of them are executing BLAST application for e-PCR: PrimerMaster (Proutski and Holmes, 1996), PRIMO (Li *et al.*, 1997), PRIMER3 (Rozen and Skaletsky, 2000), MEDUSA (Podowski and Sonnhammer, 2001), PrimeArray (Raddatz *et al.*, 2001), GST-PRIME (Varotto *et al.*, 2001), PIRA-PCR (Ke *et al.*, 2001), OligoArray (Rouillard *et al.*, 2002), PRIMEGENS (Xu *et al.*, 2002), GenomePrimer (van Hijum *et al.*, 2003), GenomePRIDE (Haas *et al.*, 2003), PUNS (Boutros and Okey, 2004), ROSO (Reymond *et al.*, 2004), GenoFrag (Ben Zakour *et al.*, 2004), MPrime (Rouchka *et al.*, 2005), SNPbox (Weckx *et al.*, 2005), Primaclade (Gadberry *et al.*, 2005), DualPrime (Andersson *et al.*, 2005) and FastPCR (http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm). However, the low speeds of BLAST or inability to process large genome sizes are the bottlenecks for these applications.

The high-speed methods applicable to large-scale projects are becoming more important with the increasing number of available full genome sequences. To overcome that problem alternative sequence search and alignment methods are required. MEGABLAST (Zhang *et al.*, 2000) is the upgrade of BLAST that is specifically designed to search highly similar matches. It is using a greedy algorithm when extending the alignment diagonals and achieves 10 times faster execution times than BLAST. MPBLAST (Korf and Gish, 2000) is a small subsidiary method that fastens the BLAST search by concatenating short query sequences into relatively few long sequences. The maximum speed improvement is about 10-fold using MPBLAST. SSAHA (Ning *et al.*, 2001) and

BLAT (Kent, 2002) are both indexing the sequence database in a similar way. Both programs build up and index of n-mers and their positions in the database. Unlike SSAHA that is using always a single perfect match as a seed, BLAT implements "unsplicing" logic – a very quick algorithm for finding short multiple nearby perfect matches. Multiple nearby matches offer much greater specificity for a given level on sensitivity than the perfect matches as shown by Jim Kent (Kent, 2002). Despite of the fact that these programs are relatively fast, they are not optimized for finding short oligonucleotides and there is a need of specific parsers to interpret the output (to count the primer binding sites and predict possible products).

The e-PCR (Schuler, 1997) program is the first application specifically designed for the prediction of all possible PCR products from given genomic sequences. It is using a word-based (7 nucleotides from the primer 3' end) strategy to speed up the search process. Program also allows using mismatches, but only in 5' end of the primer sequence. This limitation is based on the assumption that the mismatches cannot be tolerated in the 3'end of the primers (Sommer and Tautz, 1989). A web-based tool VPCR (Virtual PCR) (Lexa *et al.*, 2001) processes PCR primers, obtains BLAST search results and prints out potential PCR products. PRIMEX (Lexa and Valle, 2003) is an upgrade of previous program that is using word-based lookups from pre-indexed array of n-mers instead of BLAST searches. Sven Rahmann introduces alternative method that is using a suffix tree and the longest common substring approach for selecting the candidate oligonucleotides (Rahmann, 2003). Kevin Murphy with his co-workers have modified the original e-PCR algorithm to perform more accurate and faster string searches with their new method called me-PCR (Murphy *et al.*, 2004). The upgrade includes: the increase of maximum hash word size, hash word can be any substring of a given primer (in e-PCR it was strictly at 5' end) and multithreading for computers with several CPUs. Osprey (Gordon and Sensen, 2004) is a software package for the selection of unique and optimal oligonucleotides for microarrays and DNA sequencing. The package includes a novel computational method for the identification of alternative binding sites using position-specific scoring matrices that can be used to encode the thermodynamic profile of a sequence. This methodology is advantageous over pair-wise alignment approaches because the match and mismatch scores depend on the Nearest-Neighbor (SantaLucia, 1998) thermodynamics and therefore the secondary binding site calculation is context sensitive. This allows a more detailed evaluation of primer candidates in the oligo design process. SPCR (Cao *et al.*, 2005) can assess the similarity between primer and template using the vectors of hydrogen bond numbers after sequence conversion. This similarity (or dissimilarity) between primer and template can be used as a probability estimation of annealing site selection and annealed structure stability. Additionally, SPCR algorithm tolerates any type and number of mismatches in primer-template interaction. BISEARCH (Aranyi *et al.*, 2006) is a nice and efficient web application to design PCR primers and run e-PCR. It is

using hashing of 16-mer oligonucleotides and their permutations to identify all alternative primer locations on native genomic or bisulfate treated genomic DNA. In-Silico PCR (isPCR) (http://www.soe.ucsc.edu/~kent/src/) is a great tool created by Jim Kent for predicting PCR products using UCSC Genome Browser (http://genome.ucsc.edu/) genomic data.

6

2. PRESENT INVESTIGATIONS AND DISCUSSION

2.1. Aims of the present study

The main goal of the present study was to investigate the factors affecting the PCR success and create the effective methodology for finding and masking repeats in large genomic DNA sequences.

The specific aims for the current thesis were following:

1. to create and test a fast and efficient repeat-masking methodology suitable for applications using PCR. This methodology should be usable in large- and small-scale projects, wherein researchers are amplifying regions from genomic DNA. (Ref. II, III)
2. to create a fast and brute-force method to count the binding sites of PCR primers and predict PCR products for already designed oligonucleotides. Given methodology would allow us to examine and understand the links between word-based search methodology and PCR success rate. (Ref. I, II)
3. to find main factors related to the primer sequence that allow to predict the failure rate of PCR and compare statistical models of different complexity for their ability to predict PCR failure rate in genomic DNA sequences. (Ref. IV)

2.2. GENOMEMASKER package (Ref. II)

We have got involved with the primer design problems a several years ago by participating a large-scale genotyping project covering the human chromosome 22 (Dawson *et al.*, 2002). This study included the design of specific PCR primer pairs to amplify regions around 1278 single-nucleotide polymorphisms (SNPs). We wanted to analyze given primers from that project to study the effect of the secondary binding sites to PCR reaction success amongst other primer properties. The long running times of the whole genome database searches with current applications (BLAST) or the inability to use large input size (VPCR) gave us a reason to develop our own method (GenomeTester) for counting primer binding sites and predicting products.

The alternative approach to design unique PCR primers is to pre-mask the repeats on the template DNA. There were published no such exhaustive and fast repeat-masking tools specialized for PCR primer design. The goal was not to write another detailed annotation program of repeats for genomic DNA, but to create an application capable of finding out all short oligonucleotides in given length that are present too many times in genomic sequence. By finding and marking those short sequences primer design programs can use that for

excluding non-unique primer candidates from the template sequence. Although widely used applications for masking DNA databases and genome sequences were still RepeatMasker/MaskerAid, TandemRepeatFinder and DUST, the speed or low level of sensitivity were the main drawbacks of these programs. The application called GenomeMasker is dedicated to mask repeated primer binding sites efficiently in large genomes.

2.2.1. GenomeMasker application

2.2.1.1. Algorithm

The efficiency of both parts of the GENOMEMASKER package – GenomeTester (GT) and GenomeMasker (GM) – is based on the usage of specific hash-like data structure for genomic sequences. The hash structure in GM application contains a list of all repeated sequence motifs with given length. All words (motifs) are encoded to binary form (into 32-bit integers) and sorted to speed up the search process and reduce the size of the hash structures. The word size can be defined between 8 to 16 nucleotides in current implementation (by default it is 16).

The workflow of the GM is described graphically on Figure 1A in Ref. II. The first part of the application creates list of repeated motifs and second part masks the over-represented words in input file. The motif becomes over-represented when it appears more times in given genome as special user-defined cutoff (e.g. 1, 2, 3, etc.). The search itself is based on the binary search algorithm explained briefly in Ref. II (pg. 4). The second advantage of our method is the on-demand memory-mapping technique that will help to achieve fast search times for both small and large input data (Ref. II in pg. 4).

The third part of the GM application is a modified PRIMER3 program published lately (Koressaar and Remm, 2007). The improvements include: new formulas for calculating melting temperature and a salt correction, calculation of the effects of divalent cations and the ability to recognize and use the lower-case masked sequence for primer design. The program rejects primer candidates containing lower-case letters in 3' end. The lower-case masking preserves the DNA sequence and allows primers to be designed that partly overlap the masked region.

2.2.1.2. Sensitivity and specificity

We have tested the sensitivity of the GM and compared it with widely used program RepeatMasker (RM) at similar sensitivity level. For that we have selected 1000 random regions from human genome (1000 nt each). All these sequences were masked with both programs separately. Although the sensitivity of GM and RM was very similar (37% and 41% respectively), the sequence masking with RM is less detailed (Figure 2AB in Ref. II). The reason for this is

the incompleteness of the RepBase libraries in case of short repeats. In some cases, the DNA sequences are extensively masked by RM and the primer design in these complicated regions is impossible (Figure 2A in Ref. II). The exhaustive masking with GM will find and mask all short repeated motifs where undesired primer hybridization can occur.

To compare the specificity of different masking programs we have tested several repeat-masking programs (Table 1 in Ref. II). The primers design was attempted with PRIMER3 with combination of each masking program for all those random sequences created previously. The results clearly show that neither DUST, TandemRepeatFinder nor PRIMER3 built-in repeat library are sufficient to exclude non-unique primer candidates. RM is a good method in avoiding most of the repeats, but it is too stringent on many sequences (31% of 1000 sequences are excluded). Only 7% of the sequences masked by GM are unsuitable for primer design, thus, making the GM more suitable for PCR applications.

We have studied the effect of the primers overlapping the repeat sites. So far we have believed that masking one nucleotide from 3' end of the primer candidate is enough to guarantee the unique PCR primer. So the question is, whether GM should mask the whole repeat motif or the one nucleotide from 3' end is sufficient? Additionally, does the 5' end of the primer affect the outcome of PCR when overlapping repeat motif? To ask these questions we have selected a region from human genome that contains one repetitive sub region (words occur more than 10 times in human genome) and designed several primers overlapping the flanks of this sub region (Figure 4). We have used the GM to locate that repeat region with following parameters: word size is 16 nucleotides, masking type is ‚forward' (sense strand only) and masked only one base from 3' end of each repeat motif. There are 19 different sense primers and one antisense primer in each PCR reaction: eleven primers for testing the 3' end and eight primers for 5'end theory. The PCR protocol for these experiments is described in the Ref. IV. As shown in Figure 4, masking only one nucleotide from motifs 3' end is not enough for successful PCR reaction. It is best to mask the whole word instead of fraction of that word. This can be easily achieved with GM by defining special parameter (-nbases). The other part of this experiment gave also very interesting results. It seems that we should mask not only the motif area, but some additional nucleotides after the 3' end as well. This may show that primer whose 5' end overlaps with repeat can possibly still bind to secondary sites and therefore create alternative products by itself. Currently, the masking from that direction is not implemented in GM, but will be done in future releases.

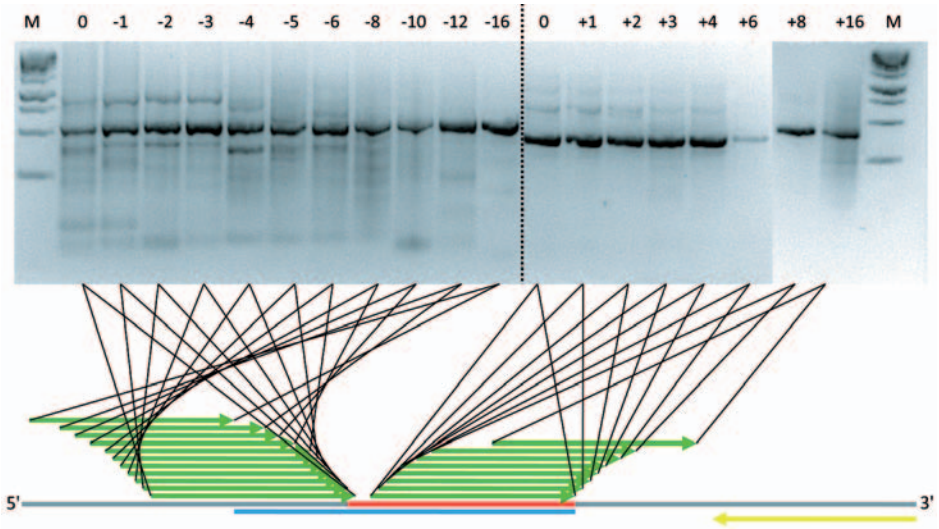


Figure 4. The effect of the repeat position in PCR primer. There are 19 different sense primers (green arrows) and on antisense primer (yellow arrow) involved with these experiments. Red line shows the repeat region masked with GenomeMasker (*wordsize=16, maskingtype=forward, nbases=1*) and blue line shows the actual repeat region. Numbers above the bands define the primer 3' end positions according to the red line. The antisense primer was the same in each reaction with different sense primer. The band with –16 above is the control experiment in which both primers are not overlapping with the actual repeat region (blue line).

2.2.1.3. Performance

One of the important aspects of evaluating program efficiency is its speed. Both methods, GM and RM, gave good results of filtering out non-unique binding sites. Therefore we decided to compare the computational performance of these two programs with different input sizes, sensitivity parameters (RM) and word lengths (GM). Even with the „rush job" (-qq) setting enabled, the RM is still at least ten times slower (Figure 3 in Ref. II). Although there is a speed-up called MaskerAid available, it makes RM even slower with the least sensitive mode than native version (Bedell *et al.*, 2000). The newer versions of RM utilize the WU-BLAST algorithm natively, without the need for MaskerAid.

2.2.2. GenomeTester application

2.2.2.1. Algorithm

The GT application counts and locates all potential binding sites of the PCR primer pair in the genome and predicts the location of all PCR products these primers can generate. The GT is based also on pre-indexed genome sequences like GM, but the main difference is in the structure of the index. Whereas GM stores only the words that are over-represented in the genome, the GT stores all locations of each word (with length defined by the user) and store them similarly to the sorted binary files. The index files (one for each letter: A, C, G, T) stores 8 bytes of data for each word occurring in genomic DNA. The first 4 bytes contain the word sequence and last four the location of the word in given chromosome/genome.

The workflow of the GT is described graphically on Figure 1B in Ref. II. Before searching primer binding sites and products with GT, one must create the binary indexes of the genomic sequences. The second part of the GT application creates a list of primer binding site coordinates and detects possible products with given length. The search itself is using the same binary search algorithm as GM to find those binding sites quickly.

2.2.2.2. Performance

The speed of the e-PCR methods working with eukaryotic genomes is the most important factor followed by memory requirement. We have created 5 different randomly selected primer datasets from human genome and tested the efficiency of several methods suitable for e-PCR (Figure 4 in Ref. II). The well-known homology search programs like BLAST and MEGABLAST are more than 100-fold slower than the newer methods. SSAHA, me-PCR and isPCR are more effective with large datasets, but GT is effective with both large and small datasets. The me-PCR is designed to predict PCR products only and in our tests, some of the products were lost with non-unique primer pairs (we were using default margin ‚M' value). The increase of this parameter will slow down the program. The memory requirements for these calculations on human chromosomes were ranging from 1 GB (SSAHA) to 300 MB (all other methods except GT) and our method was between of them allocating 500 MB of computer RAM.

2.3. Implementation for MultiPLX (Ref. I)

Large-scale studies pose the complex requirements on primer design and also on selecting primers into groups to mix them into one PCR reaction (multi-plexing). Primers in the mix must be specific to their targets and work under the

same reaction conditions. The simple string comparisons are unlikely to give accurate predictions of real interactions, therefore more advanced methods are required using Nearest-Neighbor thermodynamic alignment computation (Kaderali and Schliep, 2002). MultiPLX is designed to perform an automatic grouping of PCR primers using thermodynamic approach and can handle large datasets very efficiently. Program estimates the primer-primer and primer-product interactions, difference in T_m and product length and predicts the risk of primers generating secondary products from the template DNA. The speed of the MultiPLX algorithm is reasonable even with the larger primer sets, although the computation of primer-product interactions with very large data sets may take some time (Table 1 in Ref. I).

The calculation of the primer and product compatibility scores is implemented internally to the MultiPLX program. However, program allows the import of a custom user-specified score to help selecting optimal multiplex groups. One possibility to calculate the custom scores is to test the uniqueness of primers from different pairs that can generate alternative PCR products when multiplexed together. Therefore, we have created a special application called GT4MULTIPLX (http://bioinfo.ut.ee/gt4multiplx/), which is based on the GT algorithm described in previous chapter. It is using the similar input file as the MultiPLX (tabulated text file with ids and primer sequences) and generates all possible primer combinations of them. When GT detects one or more possible PCR products, the IDs and number of product(s) will be stored. Output of this program helps to eliminate wrong PCR product within all multiplex groups as the number of products can be thought as a specific score. User can also define a cutoff to this custom score in the MultiPLX grouping module with the parameter “maximum allowed score”.

2.4. Implementation for SNPmasker (Ref. III)

The discovery, validation and allele determination of single nucleotide polymorphisms (SNPs) can be conducted with different technologies available today (Syvanen, 2005). These methods require mostly high-quality PCR primers or probes to analyze SNPs and attention has to be paid to the repeats and variations when dealing with the genomic DNA. It is shown that the closely located SNPs are causing the lower performance on large-scale genotyping assays in the HapMap Project (Koboldt *et al.*, 2006). To overcome those problems with repeats and SNPs one should mask the template sequence before starting to design primers on it. There are several web services, which provide masking SNPs and repeats simultaneously (Table 1 in Ref. III). However, none of them allow the retrieval of masked sequence by both chromosomal coordinates and homology search. We have developed a web service called SNPmasker designed to mask SNPs from recent dbSNP database (Sherry *et al.*, 2001),

repeats with two alternative programs (GM and RM) and to offer population-specific substitution of SNP alleles using HapMap frequency tables. SNPmasker supports currently information about two organisms: human and mouse.

The implementation of SNPmasker involves three following steps: the localization of input sequences, masking of SNPs and masking of repeats. The most time consuming process is the homology search with MEGABLAST, if the exact location of the input is not defined by the user. After the sequence has been localized or retrieved from database, all SNPs (except deletions and insertions) will be masked in that region. In addition to several masking types (IUPAC,"N", custom symbol) SNPmasker provides unique option to modify the sequence by replacing SNP positions with the most frequent nucleotide (major allele) in given population (CEPH, Japanese, Chinese and African). It might be useful in studies, which are working with the individuals from specific population only as the 25% of the SNP positions (~900000 nucleotides in total) present the minor allele in the current human genomic sequence (data not shown). The masking of repeats is optional, but recommendable. There are various masking options for GM and also the possibility to use the RM (Figure 1 in Ref. III).

The masking style depends on the requirements of given study. For example, to amplify a region around SNP on could use strand-specific lower-case GM repeat-masking and replace all SNPs with "N" letter (Figure 2B in Ref. III). This kind of masking allows finding more primer candidates in highly repetitive regions. Some might want to use the RM masked sequence (e.g. for hybridization probe design) (Figure 2C in Ref. III). The usefulness of a population specific masking is already described above (Figure 2D in Ref. III).

2.5. Predicting the PCR failure rate (Ref. IV)

The statistical modeling in the field of primer design is a good possibility to estimate the weights of various molecular and sequence-specific mechanisms affecting the PCR assays. The values for these mechanisms, factors from now on, can be calculated with several software implementations available today. Given study was focused on refining the previous repeat-masking algorithm of GenomeMasker application by finding the most significant sequence-based factors causing the PCR failure.

2.5.1. Factor and model types

In this study we had the opportunity to analyze 1014 different primer pairs from human chromosome 22 (Dawson *et al.*, 2002) and 300 from random regions around the genome. For each primer pair we have selected and calculated several factors (236 in total) with various tools that may be related to the PCR failure. The important parts of the factors include different modeling of PCR primer binding sites (exact matches, mismatches, thermodynamics). There were also other primer-specific and PCR product related factors present in the statistical analysis (Table 1 in Ref. IV).

The factors are grouped differently into 5 models: GM1, GM1MM, GM2, GM2MM and PCR (Figure 1 in Ref. IV). The first four models (‚GM' can be defined as the abbreviation for GenomeMasker) contain mostly primer binding site counting properties, whereas the last model includes all factors in model building process. The binding site factors in GM1 and GM1MM models are based on the fixed word sizes (exact and with mismatches respectively) and GM2 with GM2MM on the variable word sizes (thermodynamic approach). The complexity and the computing power requirement of the parameter calculation are rising from GM1 to PCR. Although, the variety of factors is higher when building the complex models, the simpler ones are preferred in case of the similar statistical power to make the potential future implementations highly efficient.

2.5.2. Comparison of models and top factors

For each model the four most significant factors were selected and included into final models (Table 2 in Ref. IV). The statistical analysis was performed with the generalized linear models (GLZ). The order of the factors in these models is based on the χ^2 values of over the whole dataset. Interestingly, the most significant factors are the primer binding sites in each model. Other important factors include GC content of primer pairs and number of PCR products along with their length. The difference between exact and mismatched binding site modeling is minor in both, variable (GM2) and fixed (GM1) word sizes. However, comparing the first factors in each model, the thermodynamic approach gave almost two times higher χ^2 values than counting fixed strings. This confirms the arguments about better prediction of primer mispriming sites using thermodynamic modeling (SantaLucia, 2007).

The next obvious question is whether the single best factor is enough to actually eliminate the bad primer candidates or not? To answer that we have generated ten non-overlapping "control" primer sets from the original dataset to analyze the PCR failure prediction efficiency using different number of factors in each model (Figure 2 in Ref. IV). Failure rate of experimentally tested PCR pairs (predictive power of the model) was calculated at increasing sensitivity for

each model. The cutoff values are raised from 0 to given point, where the number of positive (remaining) primers is in predefined model sensitivity level (10%, 20%, 30% etc.). The simpler models, like GM1 and GM1MM, which do not include thermodynamics, were not so successful if only single factor was included into model (Figure 2A in Ref. IV). However, those models gain more power using more than one factor and reduce the difference with complex models (Figure 2B in Ref. IV). The best model GM1 helps to achieve 3-fold decrease in the failure rate of primers in our dataset: from 17% to 6%.

The binding sites with shorter word sizes and primer GC content in simple models (GM1) compensate the absence of mismatches and thermodynamics respectively. The dynamics of failure rate on some of the top factors is shown in (Figure 3 in Ref. IV). The higher number of binding sites raises the failure rate of PCR in all cases (Figure 3A in Ref. IV). High GC content in primer sequences tends to cause the PCR failure with higher probability due to possible false priming with strong energy levels in genomic DNA (Figure. 3B in Ref. IV). The higher number of PCR products (Figure. 3C in Ref. IV) increases also the failure rate, however, adding this factor to the PCR model does not make the model more efficient. Similar effect was seen with product length (fourth factor in PCR model).

2.5.3. Performance of the GM1 model

We have compared the GM1 model efficiency with widely used RepeatMasker and our previous tool GenomeMasker. For that we have selected 1000 random regions around human genome containing 1000 nucleotides each. We have masked these sequences using tools or model named above and executed PRIMER3 to design primer pairs for each region. Masking with GenomeMasker software and GM1 model is done using a special option: mask only last nucleotide from 3' end of the repeat motif. Additionally, 1000 random exonic and intronic sequences were retrieved randomly from all known human genes to compare the overall masking extent in different genomic regions.

Table 3. in Ref. IV shows that GM1 model with strict cutoff level the failure rate is approximately 2.3 times lower comparing to RepeatMasker. However, using given cutoff (10%) the primer design is possible only in 6 or 14% of the random genomic regions. Therefore, increasing the cutoff level to 20%, the primer design possibility is raising to similar level with the RepeatMasker and the failure rate of the PCR is still 1.5 times lower with GM1. Overall genome masking percentage with RepeatMasker was 50%, with GenomeMasker (max 10 binding sites allowed, masking 1 nucleotide) 52% and with GM1 model (with 4 factors, masking only one nucleotide from 3' end of the repeated word) 81% of nucleotides of human genome. Higher masking of exon regions by our GM1 method may reflect the ability of GM1 to take GC-content of primers into

account. Generally GC-rich primers have higher failure rate and therefore GC-rich exon regions are more extensively masked (Figure 3B in Ref. IV).

Although the GM1 model with four factors can reduce the PCR failure rate more than 30 percent, some of the causes of reaction failure remain still undetected. It is said that the optimization of the annealing temperature in thermocycling, salt and primer concentration, the choice of buffer and usage of enhancers can raise the good yield of unique PCR amplicon (Innis and Gelfand, 1990, Beasley *et al.*, 1999) up to 20% (SantaLucia, 2007). Therefore, the wise combination of the masking strategy with improved experimental design principles is a good way to increase the specificity and minimize the necessity of the cost- and time-expensive experimental optimization.

The results in given study demonstrate that GM1, and specifically the binding site modeling using exact matches with fixed word sizes, was similarly efficient as GM2 and PCR model and more than 2 times effective than RepeatMasker for reducing the PCR failure rate. We have compared different binding site modeling possibilities and found that the GM1 model with four factors is efficient enough to use instead of GM2MM or PCR models requiring complicated algorithmic improvements. The significant factors in GM1 model can be implemented in future versions of the GenomeMasker application and the cutoff values for word sizes should be replaced with failure rates to create even more efficient repeat-masking algorithms optimized for PCR assays.

CONCLUSIONS

The summarized results of the study:

1. We have created very fast and efficient repeat-masking and e-PCR applications included in GENOMEMASKER package. GenomeMasker application is able to mask entire human genomic DNA within 6 hours using detailed masking profile. The masking of the repeat motifs is more sensitive and specific compared to other available tools and thus being very useful in primer design assays including three main steps: masking repeats, designing primers with enhanced program modified PRIMER3 and removing primer pairs creating possible alternative PCR products. Additionally, we have created useful web interface called SNPmasker for masking repeats and SNPs with desired locations in mouse and human genomic DNA.
2. The GenomeTester application (the second important part of the GENOMEMASKER package) locates all binding sites and predicts PCR products with the speed of 1000 primer pairs per minute. The speed of the given application allowed us to create a special procedure for MULTIPLX program called GT4MULTIPLX. This allows user to calculate specific scores for MULTIPLX to achieve more accurate and successful grouping of primer pairs for multiplex PCR. In addition, the fast GenomeTester application made possible in following study to count primer binding sites with different word sizes in reasonable time-scale.
3. The statistical analysis of 236 factors that may affect the outcome of PCR reaction was performed on 1314 primer pairs and their product sequences. The most significant factors in each model we have created in this study were connected to counting primer binding sites. Additionally, the GC content of primer 3' terminus was important factor to increase the power of simpler models. The best model to use in repeat-masking applications should be as effective and easy to compute as possible. We have found that the GM1 model with four factors was similarly effective for predicting the PCR failure rate as more complex models and comparable even with PCR model, which was built including all factors used in study. These results allow us to enhance the future versions of GenomeMasker application and increase the performance of pre-masking repeats even further used in primer design process.

REFERENCES

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215,** 403–410.
2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25,** 3389–3402.
3. Andersson, A., Bernander, R. and Nilsson, P. (2005) Dual-genome primer design for construction of DNA microarrays. *Bioinformatics*, **21,** 325–332.
4. Aranyi, T., Varadi, A., Simon, I. and Tusnady, G.E. (2006) The BiSearch web server. *BMC Bioinformatics*, **7,** 431.
5. Bao, Z. and Eddy, S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, **12,** 1269–1276.
6. Beasley, E.M., Myers, R.M., Cox, D.R. and Lazzeroni, L.C. (1999) In Innis, M. A., D.H., G. and Sninsky, J. J. (eds.), *PCR Applications: Protocols for Functional Genomics*. Academic Press, San Diego, California, USA, pp. 55–72.
7. Bedell, J.A., Korf, I. and Gish, W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16,** 1040–1041.
8. Ben Zakour, N., Gautier, M., Andonov, R., Lavenier, D., Cochet, M.F., Veber, P., Sorokin, A. and Le Loir, Y. (2004) GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification. *Nucleic Acids Res*, **32,** 17–24.
9. Benita, Y., Oosting, R.S., Lok, M.C., Wise, M.J. and Humphery-Smith, I. (2003) Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res*, **31,** e99.
10. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27,** 573–580.
11. Boutros, P.C. and Okey, A.B. (2004) PUNS: transcriptomic- and genomic-in silico PCR for enhanced primer design. *Bioinformatics*, **20,** 2399–2400.
12. Budowle, B., Bieber, F.R. and Eisenberg, A.J. (2005) Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Leg Med (Tokyo)*, **7,** 230–243.
13. Campagna, D., Romualdi, C., Vitulo, N., Del Favero, M., Lexa, M., Cannata, N. and Valle, G. (2005) RAP: a new computer program for de novo identification of repeated sequences in whole genomes. *Bioinformatics*, **21,** 582–588.
14. Cao, Y., Wang, L., Xu, K., Kou, C., Zhang, Y., Wei, G., He, J., Wang, Y. and Zhao, L. (2005) Information theory-based algorithm for in silico prediction of PCR products with whole genomic sequences as templates. *BMC Bioinformatics*, **6,** 190.
15. Chavali, S., Mahajan, A., Tabassum, R., Maiti, S. and Bharadwaj, D. (2005) Oligonucleotide properties determination and primer designing: a critical examination of predictions. *Bioinformatics*, **21,** 3918–3925.
16. Chen, S.H., Lin, C.Y., Cho, C.S., Lo, C.Z. and Hsiung, C.A. (2003) Primer Design Assistant (PDA): A web-based primer design tool. *Nucleic Acids Res*, **31,** 3751–3754.
17. Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418,** 544–548.

9

18. Deininger, P.L. and Batzer, M.A. (1999) Alu repeats and human disease. *Mol Genet Metab*, **67,** 183–193.
19. Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res*, **27,** 2369–2376.
20. Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res*, **12,** 387–395.
21. Dong, F., Allawi, H.T., Anderson, T., Neri, B.P. and Lyamichev, V.I. (2001) Secondary structure prediction and structure-specific sequence analysis of single-stranded DNA. *Nucleic Acids Res*, **29,** 3248–3257.
22. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32,** 1792–1797.
23. Edgar, R.C. and Myers, E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21 Suppl 1,** i152–158.
24. Erlich, H.A., Gelfand, D. and Sninsky, J.J. (1991) Recent advances in the polymerase chain reaction. *Science*, **252,** 1643–1651.
25. Fedorova, O.S., Podust, L.M., Maksakova, G.A., Gorn, V.V. and Knorre, D.G. (1992) The influence of the target structure on the efficiency of alkylation of single-stranded DNA with the reactive derivatives of antisense oligonucleotides. *FEBS Lett*, **302,** 47–50.
26. Gadberry, M.D., Malcomber, S.T., Doust, A.N. and Kellogg, E.A. (2005) Primaclade – a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, **21,** 1263–1264.
27. Gordon, P.M. and Sensen, C.W. (2004) Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays. *Nucleic Acids Res*, **32,** e133.
28. Guyer, R.L. and Koshland, D.E., Jr. (1989) The Molecule of the Year. *Science*, **246,** 1543–1546.
29. Haas, S., Vingron, M., Poustka, A. and Wiemann, S. (1998) Primer design for large scale sequencing. *Nucleic Acids Res*, **26,** 3006–3012.
30. Haas, S.A., Hild, M., Wright, A.P., Hain, T., Talibi, D. and Vingron, M. (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res*, **31,** 5576–5581.
31. Heid, C.A., Stevens, J., Livak, K.J. and Williams, P.M. (1996) Real time quantitative PCR. *Genome Res*, **6,** 986–994.
32. Higuchi, R., Fockler, C., Dollinger, G. and Watson, R. (1993) Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology (N Y)*, **11,** 1026–1030.
33. Housley, D.J., Zalewski, Z.A., Beckett, S.E. and Venta, P.J. (2006) Design factors that influence PCR amplification success of cross-species primers among 1147 mammalian primer pairs. *BMC Genomics*, **7,** 253.
34. Innis, M.A. and Gelfand, D.H. (1990) In Innis, Gelfand, Sninsky and White (eds.), *PCR Protocols*. Academic Press, New York, pp. 3–12.
35. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, **16,** 418–420.
36. Kaderali, L. and Schliep, A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18,** 1340–1349.
37. Kazazian, H.H., Jr. (2004) Mobile elements: drivers of genome evolution. *Science*, **303,** 1626–1632.

38. Ke, X., Collins, A. and Ye, S. (2001) PIRA PCR designer for restriction analysis of single nucleotide polymorphisms. *Bioinformatics*, **17,** 838–839.
39. Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res*, **12,** 656–664.
40. Koboldt, D.C., Miller, R.D. and Kwok, P.Y. (2006) Distribution of human SNPs and its effect on high-throughput genotyping. *Hum Mutat*, **27,** 249–254.
41. Kohany, O., Gentles, A.J., Hankus, L. and Jurka, J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7,** 474.
42. Koressaar, T. and Remm, M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23,** 1289–1291.
43. Korf, I. and Gish, W. (2000) MPBLAST : improved BLAST performance with multiplexed queries. *Bioinformatics*, **16,** 1052–1053.
44. Kreil, D.P., Russell, R.R. and Russell, S. (2006) Microarray oligonucleotide probes. *Methods Enzymol*, **410,** 73–98.
45. Kubista, M., Andrade, J.M., Bengtsson, M., Forootan, A., Jonak, J., Lind, K., Sindelka, R., Sjoback, R., Sjogreen, B., Strombom, L. *et al.* (2006) The real-time polymerase chain reaction. *Mol Aspects Med*, **27,** 95–125.
46. Kurg, A., Tonisson, N., Georgiou, I., Shumaker, J., Tollett, J. and Metspalu, A. (2000) Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet Test*, **4,** 1–7.
47. Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, **29,** 4633–4642.
48. Kurtz, S. and Schleiermacher, C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15,** 426–427.
49. Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J.J. (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Res*, **18,** 999–1005.
50. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409,** 860–921.
51. Lawyer, F.C., Stoffel, S., Saiki, R.K., Myambo, K., Drummond, R. and Gelfand, D.H. (1989) Isolation, characterization, and expression in Escherichia coli of the DNA polymerase gene from Thermus aquaticus. *J Biol Chem*, **264,** 6427–6437.
52. Lefebvre, A., Lecroq, T., Dauchel, H. and Alexandre, J. (2003) FORRepeats: detects repeats on entire chromosomes and between genomes. *Bioinformatics*, **19,** 319–326.
53. Lexa, M., Horak, J. and Brzobohaty, B. (2001) Virtual PCR. *Bioinformatics*, **17,** 192–193.
54. Lexa, M. and Valle, G. (2003) PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*, **19,** 2486–2488.
55. Li, P., Kupfer, K.C., Davies, C.J., Burbee, D., Evans, G.A. and Garner, H.R. (1997) PRIMO: A primer design program that applies base quality statistics for automated large-scale DNA sequencing. *Genomics*, **40,** 476–485.
56. Li, X., Kahveci, T. and Settles, A.M. (2008) A novel genome-scale repeat finder geared towards transposons. *Bioinformatics*, **24,** 468–476.

57. Martinez, H.M. (1983) An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res*, **11,** 4629–4634.
58. Miura, F., Uematsu, C., Sakaki, Y. and Ito, T. (2005) A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. *Bioinformatics*, **21,** 4363–4370.
59. Morgulis, A., Gertz, E.M., Schaffer, A.A. and Agarwala, R. (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, **22,** 134–141.
60. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol*, **51 Pt 1,** 263–273.
61. Mullis, K.B. and Faloona, F.A. (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*, **155,** 335–350.
62. Murphy, K., Raj, T., Winters, R.S. and White, P.S. (2004) me-PCR: a refined ultrafast algorithm for identifying sequence-defined genomic elements. *Bioinformatics*, **20,** 588–590.
63. Nagashima, T., Matsuda, H., Silva, D.G., Petrovsky, N., Konagaya, A., Schonbach, C., Kasukawa, T., Arakawa, T., Carninci, P., Kawai, J. *et al.* (2004) FREP: a database of functional repeats in mouse cDNAs. *Nucleic Acids Res*, **32,** D471–475.
64. Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res*, **11,** 1725–1729.
65. Nugent, K.G. and Saville, B.J. (2004) Forensic analysis of hallucinogenic fungi: a DNA-based approach. *Forensic Sci Int*, **140,** 147–157.
66. Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) A common language for physical mapping of the human genome. *Science*, **245,** 1434–1435.
67. Onodera, K. and Melcher, U. (2004) Selection for 3' end triplets for polymerase chain reaction primers. *Mol Cell Probes*, **18,** 369–372.
68. Owczarzy, R., Vallone, P.M., Gallo, F.J., Paner, T.M., Lane, M.J. and Benight, A.S. (1997) Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers*, **44,** 217–239.
69. Panjkovich, A. and Melo, F. (2005) Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics*, **21,** 711–722.
70. Pevzner, P.A., Tang, H. and Tesler, G. (2004) De novo repeat classification and fragment assembly. *Genome Res*, **14,** 1786–1796.
71. Podowski, R.M. and Sonnhammer, E.L. (2001) MEDUSA: large scale automatic selection and visual assessment of PCR primer pairs. *Bioinformatics*, **17,** 656–657.
72. Prak, E.T. and Kazazian, H.H., Jr. (2000) Mobile elements and the human genome. *Nat Rev Genet*, **1,** 134–144.
73. Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21 Suppl 1,** i351–358.
74. Proutski, V. and Holmes, E.C. (1996) Primer Master: a new program for the design and analysis of PCR primers. *Comput Appl Biosci*, **12,** 253–255.
75. Raddatz, G., Dehio, M., Meyer, T.F. and Dehio, C. (2001) PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics*, **17,** 98–99.
76. Rahmann, S. (2003) Fast large scale oligonucleotide selection using the longest common factor approach. *J Bioinform Comput Biol*, **1,** 343–361.

77. Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G. and Fayard, J.M. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20,** 271–273.
78. Richards, R.I. and Sutherland, G.R. (1994) Simple repeat DNA is not replicated simply. *Nat Genet*, **6,** 114–116.
79. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, **132,** 365–386.
80. Rouchka, E.C., Khalyfa, A. and Cooper, N.G. (2005) MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays. *BMC Bioinformatics*, **6,** 175.
81. Rouillard, J.M., Herbert, C.J. and Zuker, M. (2002) OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, **18,** 486–487.
82. Rubin, E. and Levy, A.A. (1996) A mathematical model and a computerized simulation of PCR using complex templates. *Nucleic Acids Res*, **24,** 3538–3545.
83. Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1986) Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature*, **324,** 163–166.
84. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239,** 487–491.
85. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230,** 1350–1354.
86. SantaLucia, J., Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A*, **95,** 1460–1465.
87. SantaLucia, J., Jr. (2007) Physical principles and visual-OMP software for optimal PCR design. *Methods Mol Biol*, **402,** 3–34.
88. SantaLucia, J., Jr. and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct*, **33,** 415–440.
89. Scharf, S.J., Horn, G.T. and Erlich, H.A. (1986) Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*, **233,** 1076–1078.
90. Schuler, G.D. (1997) Sequence mapping by electronic PCR. *Genome Res*, **7,** 541–550.
91. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29,** 308–311.
92. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147,** 195–197.
93. Sommer, R. and Tautz, D. (1989) Minimal homology requirements for PCR primers. *Nucleic Acids Res*, **17,** 6749.
94. Syvanen, A.C. (2005) Toward genome-wide SNP genotyping. *Nat Genet*, **37 Suppl,** S5–10.
95. Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol*, **7,** 203–214.
96. Zhi, D., Raphael, B.J., Price, A.L., Tang, H. and Pevzner, P.A. (2006) Identifying repeat domains in large genomes. *Genome Biol*, **7,** R7.

10

97. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31,** 3406–3415.
98. van Hijum, S.A., de Jong, A., Buist, G., Kok, J. and Kuipers, O.P. (2003) UniFrag and GenomePrimer: selection of primers for genome-wide production of unique amplicons. *Bioinformatics*, **19,** 1580–1582.
99. Varadaraj, K. and Skinner, D.M. (1994) Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene*, **140,** 1–5.
100. Varotto, C., Richly, E., Salamini, F. and Leister, D. (2001) GST-PRIME: a genome-wide primer design software for the generation of gene sequence tags. *Nucleic Acids Res*, **29,** 4373–4377.
101. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420,** 520–562.
102. Weckx, S., De Rijk, P., Van Broeckhoven, C. and Del-Favero, J. (2005) SNPbox: a modular software package for large-scale primer design. *Bioinformatics*, **21,** 385–387.
103. Volfovsky, N., Haas, B.J. and Salzberg, S.L. (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biol*, **2,** RESEARCH0027.
104. Vollenhofer, S., Burg, K., Schmidt, J. and Kroath, H. (1999) Genetically modified organisms in food-screening and specific detection by polymerase chain reaction. *J Agric Food Chem*, **47,** 5038–5043.
105. von Ahsen, N., Wittwer, C.T. and Schutz, E. (2001) Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin Chem*, **47,** 1956–1961.
106. Xu, D., Li, G., Wu, L., Zhou, J. and Xu, Y. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, **18,** 1432–1437.
107. Yancy, H.F., Mohla, A., Farrell, D.E. and Myers, M.J. (2005) Evaluation of a rapid PCR-based method for the detection of animal material. *J Food Prot*, **68,** 2651–2655.
108. Yap, E.P. and McGee, J.O. (1991) Short PCR product yields improved by lower denaturation temperatures. *Nucleic Acids Res*, **19,** 1713.
109. Yuryev, A., Huang, J., Pohl, M., Patch, R., Watson, F., Bell, P., Donaldson, M., Phillips, M.S. and Boyce-Jacino, M.T. (2002) Predicting the success of primer extension genotyping assays using statistical modeling. *Nucleic Acids Res*, **30,** e131.

SUMMARY IN ESTONIAN

Meetodid ja tarkvara PCR praimerite töötamise ennustamiseks suurtes genoomsetes DNA järjestustes

DNA oligonukleotiididel põhinevad tehnoloogiad on leidnud biotehnoloogia valdkonnas laialdast kasutust. Üheks levinumaks molekulaarseks meetodiks on DNA polümeraasi ahelreaktsioon (PCR). Tegemist on tsüklilise reaktsiooniga, kus mõlema DNA ahela jaoks sünteesitakse protsessi käigus uus komplementaarne ahel. Lisaks reagentidele, ensüümile ja paljundatavale DNA-le, on protsessi jaoks vajalikud lühikesed oligonukleotiidid ehk PCR praimerid, mis hübridiseeruvad vastavalt komplementaarse DNA ahelaga ja võimaldavad ensüümil pikendada antud puuduvat ahelat. Ideaalseks reaktsiooni tulemuseks on spetsiifiline ja kõrge kontsentratsiooniga paljundatud DNA regioon ehk PCR produkt (Saiki *et al.*, 1988).

Kaasaegsetes suuremahulistes genotüpiseerimise projektides disainitakse ja kasutatakse tuhandeid praimeripaare korraga, et üles amplifitseerida erinevaid regioone iga indiviidi või organismi DNA pealt. Seetõttu on PCR edukust mõjutavate faktorite hindamine praimerite valimise protsessis väga oluline, et vähendada rahalisi kulutusi ja ajakulu. Varasemad uuringud selles vallas on keskendunud rohkem reaktsiooni reagentide optimiseerimisele nagu PCR puhvri komponentide, soola, DNA, oligonukleotiidide jt. kontsentratioonid ning protokolli optimiseerimisele (Innis and Gelfand, 1990, Beasley *et al.*, 1999). Hilisemad uuringud on pööranud tähelepanu ka praimerite järjestuse omadustele nagu GC sisaldus, praimeri pikkus ja sekundaarstruktuurid, mis võivad mõjutada amplifiseerimise efektiivsust (Haas *et al.*, 1998, Rozen and Skaletsky, 2000, Chen *et al.*, 2003, Chavali *et al.*, 2005, Miura *et al.*, 2005). Samuti on uuritud kindlate nukleotiidide või nende kombinatsioonide mõju praimeri erinevates positsioonides (Yuryev *et al.*, 2002) ja PCR produktide järjestusepõhiseid omadusi: GC sisaldus, sekundaarsturktuurid (Varadaraj and Skinner, 1994, Benita *et al.*, 2003). Vähem on uuritud eukarüootsetes organismides leiduvate korduvate motiivide mõju PCR edukusele.

Käesoleva doktoritöö kirjanduse ülevaade keskendub seni teadaolevate PCR reaktsiooni mõjutavate faktorite kirjeldamisele erinevate uurimisgruppide poolt. Eraldi on välja toodud biokeemilised ja järjestusepõhised faktorid. Lisaks on antud lühiülevaade eukarüootide kordusjärjestustest ja nende klassifikatsioonist ning korduste leidmise meetoditest. Viimane peatükk kirjeldab e-PCR metoodikat ja selle rakendusi, mis on tänapäeval kasutuses.

Antud doktoritöö üheks eesmärgiks oli luua kiire ja efektiivne korduste maskeerimise meetoodika, mis on spetsiaalselt optimiseeritud PCR praimerite disainiks. Doktoritöö raames loodi programmide pakett GENOMEMASKER, milles leiduv aplikatsioon GenomeMasker on võimeline maskeerima kõik korduvad motiivid inimese genoomsel DNA-l 6 tunniga. Järjestuste maskeeri-

mine on võrreldes teiste olemasolevate programmidega tunduvalt kiirem, täpsem ja spetsiifilisem. Lisaks on loodud spetsiaalne web'i aplikatsioon SNPmasker, mille abil on kasutajal võimalik maskeerida ära kordused ja ühenukleotiidsed polümorfismid (SNP) nii inimese kui hiire järjestustel. Pakett sisaldab ka modifitseeritud praimerite disaini programmi PRIMER3, mis tunneb ära GenomeMasker poolt maskeeritud DNA järjestuse ja kasutab uuemaid termodünaamika tabeleid ning valemeid.

Teiseks eesmärgiks oli luua meetod, mis võimaldaks kiiresti lugeda kokku PCR praimerite seondumiskohad suurtes genoomides ja ennustada produktide teket. GenomeTester nimeline aplikatsioon GENOMEMASKER paketis võimaldab PCR produkte ennustada 1000 praimeripaari jaoks minutis. Lähtudes aplikatsiooni kiirusest oli võimalik MULTIPLX programmi jaoks kirjutada spetsiaalne web'i tööriist GT4MULTIPLX, mis võimaldab arvutada praimeripaaride jaoks spetsiifilised skoorid, mida on võimalik hilisemal multipleks gruppide moodustamisel arvesse võtta. Lisaks võimaldas efektiivne GenomeTester aplikatsioon järgmises uuringus mõistliku aja jooksul läbi viia erinevate sõnapikkustega praimerite seondumiskohtade ja produktide kokkulugemised.

Käesoleva töö viimases osas uuriti erinevaid faktoreid, mis võiksid vähendada PCR edukust. Uuringus kasutati 1314 praimeripaari katseandmeid (>80000 üksikut katset) ja iga paari kohta arvutati 236 erineva faktori väärtused. Selgus, et kõige enam mõjutab PCR edukust praimeri seondumiskohtade arv genoomis. Lisaks oli oluline primeri 3' otsa GC sisaldus. Oluliste faktorite põhjal koostati 5 erinevat PCR edukust ennustavat statistilist mudelit. Faktorid jaotati mudelitesse arvutusliku keerukuse alusel. Mudelite võrdlemisel selgus, et GM1 (kõige lihtsamini arvutavav mudel), mis sisaldab 4 olulisemat faktorit, ennustab PCR edukust samal tasemel või isegi paremini, kui keerulisemad mudelid (GM2, PCR). Sellele toetudes on võimalik tulevikus tõsta GenomeMasker aplikatsiooni algoritmi efektiivsust veelgi praimerite disaini protsessi parandamiseks.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Prof. Maido Remm who has guided me through the whole M.Sc and Ph.D studies and who was always supportive and injected me optimism during my work. He helped me to stay focused, gave valuable guidelines and encouraged to concentrate on what is important.

I am grateful to Prof. Jaak Vilo for introducing me the bioinformatics world and giving shelter during my studies in European Bioinformatics Institute, Hinxton, UK.

I would like to acknowledge my first supervisor Prof. Andres Metspalu from the Biotechnology Department, University of Tartu. I am grateful for the opportunity to work in his lab and for the possibility to do a lot of wet-lab experiments. I am thankful to all my previous supervisors in his lab – Krista Kaasik, Maris Teder-Laving and Hardo Lilleväli – who taught me to conduct the molecular laboratory methods.

I would like to thank all the co-authors of the papers which this dissertation is based on.

I wish to thank all my former colleagues from Department of Biotechnology, Asper Biotech and Biodata for all the help and for creating friendly atmosphere. I would like to specially thank my present colleagues from Department of Bioinformatics for broadening my knowledge of Oriental and Islamic cultures, growing plants, military weapon systems, where to invest and what is really happening in the world politics.

I would like to thank my close friends: Tõnis, Viljo and Oliver. Those long fruitful discussions, frightful amounts of beer and overwhelming singing hours helped to keep me on track. Many thanks to all my friends for making my life more interesting and thrilling.

I owe my gratitude to all my family (here and overseas) who supported my studies on hard times and believed in me that I can actually make this through.

Last but not least, I would like to thank my dear wife Helena for love, tender and happiness. You have given me something that nobody could before you: two loveliest kids Grete and Oliver. Also, the never-ending support and warmth kept me going throughout these joyful years of my life.

11

PUBLICATIONS

I

Kaplinski L, **Andreson R**, Puurand T, Remm M (2005). MultiPLX: automatic grouping and evaluation of PCR primers. Bioinformatics 21(8): 1701–2.

II

Andreson R, Reppo E, Kaplinski L, Remm M (2006).
GENOMEMASKER package for designing unique genomic PCR primers.
BMC Bioinformatics 7:172.

III

Andreson R, Puurand T, Remm M (2006). SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. Nucleic Acids Research 34:W651–5.

IV

Andreson R, Möls T, Remm M (2008).
Predicting failure rate of PCR in large genomes.
Nucleic Acids Research (accepted)

Predicting failure rate of PCR in large genomes

Reidar Andreson[1,2], Tõnu Möls[1], Maido Remm[1,2*]

[1] Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Estonia
[2] Estonian Biocentre, Tartu, Estonia

* Corresponding author

Contact:
Prof. Maido Remm
Riia str. 23, Tartu 51010, ESTONIA
tel.: +372–7375001
fax: +372–7420286
e-mail: maido.remm@ut.ee

ABSTRACT

We have developed statistical models for estimating the failure rate of PCR primers using 236 primer sequence related factors. The model involved 1314 primer pairs and is based on more than 80,000 PCR experiments. We found that the most important factor in determining PCR failure is the number of predicted primer binding sites in the genomic DNA. We also compared different ways of defining primer binding sites (fixed length word vs. thermodynamic model; exact match vs. matches including 1–2 mismatches). We found that the most efficient prediction of PCR failure rates can be achieved using a combination of 4 factors (number of primer binding sites counted in different ways plus GC% of the primer) combined into single statistical model GM1. According to our estimations from experimental data, the GM1 model can reduce the average failure rate of PCR primers nearly 3-fold (from 17% to 6%). The GM1 model can easily be implemented in software to pre-mask genome sequences for potentially failing PCR primers, thus improving large-scale PCR primer design.

19

INTRODUCTION

During recent decades, the Polymerase Chain Reaction (PCR) has become a very widely-used method, routinely performed in various molecular biology applications. Ideally, there should be a pair of unique primers that amplify the desired target sequence selectively with maximum yield. A common way to deal with PCR problems is to optimize the conditions, such as the concentrations of reagents in the PCR buffer, or to modify the primer annealing temperature (1). In addition, careful design of primers is crucial for the success of PCR. The list of factors that may influence the success of PCR is long and has been studied by several investigators: the concentrations of the PCR buffer reagents, the primer length and the GC contents of primers and template, simple repeats in the primer sequence, stable secondary structures of both primer and product sequences, etc. (2–5). However, the circumstances that affect PCR and the molecular mechanisms behind it have become considerably better understood over the years. The most recent studies utilizing PCR incorporate more factors in primer design; for instance, the nucleotide composition of the primer 3' end and the size of that important region, the nucleotide composition of the PCR product, the secondary structure of the PCR template around the primer annealing sites, the T_m of the product and the regionalized GC content of the PCR template (6–8). In addition, some studies have focused specifically on certain classes of factors such as the stability and the uniqueness of the primer 3' end (9,10) or have concentrated on comparing the software available for PCR primer design (11).

Selection of non-optimal primers can lead to amplification of undesired regions or no amplification at all. Careful choice of primers is even more crucial in high-throughput assays, where the total cost of the primers is high. Previously, we have worked on algorithms for masking repeats and from these we have developed a fast algorithm for finding and masking short repeats in the human genome (12). Following this earlier work, our aim here was to carry out a comprehensive statistical analysis of potential descriptive factors that could influence the failure rates of PCR primers. We prefer to use term "failure rate" instead of "success rate" because only failures of PCR can be directly predicted from primer sequences. Success of PCR is dependent on many other factors besides primer properties (reagents, equipment, human errors, etc.) and therefore inherently less predictable.

The typical factors influencing the PCR failure rate include purity of genomic DNA, the length of DNA fragments, the precision of DNA concentration measurements, the presence of SNPs and the uniqueness of PCR primer sequences. In this paper, we present a study of the most significant primer and product sequence-specific factors in reducing the PCR failure rate to offer scenarios for improving masking algorithms for the regions in genomes that might have high PCR failure rates.

MATERIALS AND METHODS

Computation

All programs described in this article were executed on a 2.66 GHz Intel Xeon™ processor machine with 6 GB of RAM. For statistical analysis and modeling, SAS software ver. 9.1.3 (SAS OnlineDoc® 9.1.3**.,** SAS Institute Inc. 2004, Cary, NC, USA) was used. All calculations were performed on assembled chromosome sequences derived from NCBI build 35 (13). Single nucleotide polymorphism (SNP) data were retrieved from dbSNP build 125. Human exon and intron sequences were retrieved from UCSC Genome Browser (hg18) using Table Browser tool (14).

The number of exact-match binding sites for the PCR primers was calculated by the GenomeTester program from the GENOMEMASKER software package (12). The number of binding sites including mismatches was calculated using the BLAST program (15) with filtering turned off (-F F). The number of primer binding sites (exact binding or binding with mismatches) based on thermodynamic affinity was found by the FASTAGREP program (executable available from http://bioinfo.ebc.ee/download/), which was executed with default parameter values.

To mask PCR product sequences, GenomeMasker, DUST (ftp://ftp.ncbi.nlm.nih.gov/pub/tatusov/dust/) and RepeatMasker (Smit, A.F.A., Hubley, R. and Green, P. http://www.repeatmasker.org/) were used. With RepeatMasker, the RepBase Update (16) 8.12 library of repeated motifs in the human genome was used. RepeatMasker was executed with the following sensitivity parameters: –s, –q and –qq. DUST was used with default parameters. The masking program GenomeMasker was used with masking letter parameter 'l' (lower-case masking) and masking type parameter 'target 500 501'.

The melting temperature (T_m) of each primer was calculated using current thermodynamic tables (17). The secondary structures of PCR primers and products were calculated using MFOLD 3.2 (18) and primer-primer dimer bindings were calculated with MULTIPLX (19). Three additional factors (PROD_AUCGC, PROD_AUCGC2 and PROD_RATIOGC_100) for PCR products were calculated using software published by Benita and co-workers (7).

Experimental datasets

Experimental PCR data originated from two large-scale datasets. One was obtained from analysis of 1278 SNPs on human chromosome 22 (20). In our analysis, we used 1014 of these 1278 primer pairs, most of which were designed with Primer3 software (21). Approximately 20% of the primer sequences were retrieved from the scientific literature or were designed manually. In the second dataset, we had 300 primer pairs designed by Primer3 from randomly selected regions of the human genomic DNA and we performed 10 experiments with each primer pair. Both primer sets were non-redundant; all primer sequences

were unrelated (unique and non-overlapping). Primer3 was used with no repeat libraries. No other repeat detection or masking procedures were used in the primer design.

PCR primers in these datasets had the following properties. Melting temperatures varied between 48°C and 75°C, with an average of 60°C. GC contents varied between 27% and 80%, with an average of 56%. Primer lengths varied between 18 and 25 nucleotides (average 21 bp) and PCR product size was 100 to 600 bp.

Experimental conditions and testing

PCR conditions for amplifying products were as follows: 15 min pre-incubation at 95°C, followed by seven touchdown cycles of 20 s at 95°C, 30 s at 66°C (decreasing 1°C per cycle), 30 s at 72°C; seventeen cycles of 20 s at 95°C, 30 s at 58°C, 30 s at 72°C; sixteen cycles of 20 s at 95°C, 30 s at 56°C, 30 s at 72°C; and final extension at 72°C for 7 min. DNA was extracted from human blood cells by a modified salting out method (22). Each PCR reaction contained 15 ng of genomic DNA in a volume of 10 μl. All PCR reactions were conducted by Asper Biotech Ltd.

In both sets, each primer pair was tested experimentally at least 10 times (84,142 reactions in total) using DNAs from different individuals. The PCR primer failure rate was estimated by analyzing agarose gel electropherograms. Reactions were counted as positive if they gave clear bands of the expected size. Reactions with smears and multiple bands were counted as negative. The average PCR failure rate for a given PCR primer pair was expressed as the fraction of negative results.

Statistical methods and development of models

The Generalized Linear Model (GLZ) is a generalization of the general linear model (GLM) that can be used to predict responses both for dependent variables with discrete distributions and for dependent variables that are nonlinearly related to the predictors (23). The procedure of the SAS software GENMOD is based on the theory of generalized linear models. In the GENMOD procedure, both the number of successful trials and the total number of events were used, assuming a binomial distribution and logit link function.

To maximize the usage of limited numbers of primers we joined our two datasets into one and thereafter divided the dataset randomly into 10 non-overlapping segments. In such a sampling procedure, one single segment can be treated as a ”control set”. For each control set, a ”training set” was created from the remaining primer pairs. Ten different models were developed for each training set and tested against the given control set. The significance of the descriptive factors in assessing the failure of PCR was measured by χ^2 and by the corresponding *p*-values. Only factors that were statistically significant in both the training and control sets were considered further in the composition of

the final models. Factors were included in the final model building if they were statistically significant in at least five sub-models out of ten. The *type1* option of the GENMOD procedure (Type I) was used to perform forward stepwise analysis (a cutoff value of $p<0.0001$ was used) to build the final models. The values of the intercept and regression coefficients were retrieved from the models of training sets. Probabilities of failure (P) were predicted for each primer pair of the control set on the basis of the parameter estimates retrieved. The formula used by GENMOD procedure for calculating the probability P of a failure was

$$P = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}},$$

where α is the intercept (constant), β_i are regression coefficients (estimates), X_i are the independent variables (factor values) and k denotes the number of factors used in the given model. The decrease of experimental failure rates in the datasets tested was used as a measure for comparing different models.

RESULTS

The major aims of the current study were: (a) to find the main factors related to the primer sequence that allow predict the PCR failure rate; (b) to compare statistical models of different complexities for their ability to predict the failure rate of the PCR reaction conducted with genomic DNA sequences. Models differing in complexity were compared because of the need for a model for large-scale applications, such as masking the low-success rate regions in the entire human genome. Therefore, the resulting algorithm should be able to calculate the failure rate of every possible PCR primer in the eukaryotic genome in reasonable time.

Factors used to predict PCR failure rate

Two hundred and thirty six different factors describing primer or primer pair properties were selected for statistical analysis (Table 1). The values of these factors were calculated for each primer pair and for each predicted product sequence (only one PCR product per primer pair was considered here).

Calculation of primer binding sites can be modeled in various ways as shown in Table 1: for example, using a primer sequence substring of fixed or variable length (based on the thermodynamic approach), finding matches for a string with 100% identity or with mismatches. One of the major goals of this study was to determine whether the number of binding sites of (a substring of) the primer including mismatches gives a better prediction of the PCR failure rate than models that use 100% identity for model binding. Another important

20

goal was to determine whether modeling the primer binding site thermodynamically would improve the prediction of PCR failure rate.

We also included factors that are not directly related to the parameters of the number of primer binding sites: primer length, GC% of primers, Gibbs free energy of possible secondary structures and primer-dimers and T_m of primers. We have also analyzed as factors the significance of known SNPs in primer sequences and energetically stable perfect duplexes composed of the last nine bases of the 3' end of the primer. The SNPs may cause allele-specific PCR or possible failure in primer-template hybridization.

Factors specific to primer pairs include several parameters associated with the PCR product: the number of predicted PCR products, the length of the product, GC% of the product and secondary structures within the product. Other factors under study are regionalized GC contents and the number of repeat-masked nucleotides within the template DNA counted by different methods. The nucleotide combinations of primer 3' end and the first amplicon nucleotide following immediately the primer sequence were also studied.

Five models for predicting PCR failure rate

We have created 5 different types of models based on the complexity of the calculation of factors included. Model complexity can be measured by several parameters. The most important is the time needed for computing the values of factors in a given model. The calculation of primer binding sites including mismatches takes more time than the calculation of exact matches because of the substitutions required in a primer sequence when comparing the primer with genomic DNA. For example, there are 49 different sequence variants for a word containing 16 nucleotides (word size = 16) with one mismatch at any position. Executing brute-force scanning of 16-mer oligonucleotides against a DNA template without using any heuristics increases the search time more than 40-fold. To make things more complicated, determination of the free energy levels of the oligonucleotides makes these calculations even slower. Such thermodynamic binding site modeling allows us to define a word size that is biophysically more meaningful than fixed length. A sequence containing mostly A and T nucleotides has higher free energy values than a GC rich sequence. In a PCR reaction under fixed conditions, the former sequence requires more nucleotides identical with the template to hybridize in the annealing process than the latter. Therefore, variable word length must be used for searching binding sites, and more time is required to calculate these sites from genomic DNA.

Factors included in the first four models (described below) contain the properties that can be used without prior information about the primer pairs; we can use a single primer sequence to calculate the values of these factors. Thus, they are suitable for pre-masking low success-rate regions in the genomic DNA and are therefore called GenomeMasker (GM*) models. The last model requires full sequences of both primers and the PCR product for calculating the values of

all factors; therefore it can be considered as an e-PCR model, capable of predicting the outcome of a given primer pair. The overlap of factors between the different models is shown in Figure 1. Ten factors that were used in all models are: GC content of the 3' end of a primer (8, 12 and 16 nt.) and the number of SNPs and their positions in the primer sequence (16 nt. from 3' end). Additional factors included in different models were:

GM1 model is the simplest model, containing 22 different factors. These include parameters associated with the number of exact binding sites in the human genome, modeled by fixed word length.

GM1MM model contains basically the same factors, but binding sites with one and two mismatches are also present (38 factors in total). Binding site modeling within this model is similar to many primer design programs in which primer specificity is determined using BLAST software.

GM2 model is the first of the thermodynamic models that contains binding site counts (exact matches) with variable length word sizes using three levels of Gibbs free energy: $\Delta G_{37} < -10, -15, -20$ kcal/mol (16 factors in total). Those energy levels correspond to the following average word sizes in the human genome: 10 (min=6, max=15) 14 (min=8, max=22) and 18 (min=10, max=29) nt respectively.

GM2MM model includes the counts of binding sites with variable-length words using three free energy levels and one or two mismatches (28 factors in total).

PCR model was built from all the factors examined in this study (236 in total).

Selection of the most significant factors for each model and the building of the final models were achieved by 10-fold cross-validation. We divided our combined experimental dataset into ten subsets as described in the Methods section. For each sub-model, the best factors present in at least half the cross-validated datasets were included in building the final models. Their order in the model was based on the χ^2 values over the whole dataset.

The top four major factors in each model are shown in Table 2. Including more than 4 factors in a model did not significantly improve its prediction power (see below). The first two (and thus the most important) predictors for each model are related to the maximum number of primer binding sites in the human genome (Table 2). Other factors that improve the PCR model are PCR product length and the number of predicted PCR products. The GM1, GM1MM and GM2 models also included one factor that is related to primer GC-content.

Four factors are required for robust prediction of failure rate

The next step was to compare the predictive powers of these models using cross-validation datasets. The parameter estimates for the model formulae were calculated with SAS GENMOD procedure for each model using the whole dataset. The predictive power of each model was then measured and compared on the basis of the average fraction of failed primer pairs in those ten control

sets. It is always computationally easier to implement models with fewer factors. Therefore, we created graphs for each model with one to four factors to determine whether the additional factor improves the prediction of the model. The average failure rate of experimentally tested PCR pairs (predictive power of the model) was calculated over all control sets with increasing sensitivity for each model (Figure 2). Sensitivity was defined as the percentage of primer pairs remaining after the failure-prediction model was applied to our cross-validation datasets. The cutoff values were raised from 0 to the given point, where the number of positive (remaining) primer pairs is at a predefined level (10%, 20%, 30% etc.).

The predicted failure rate can be approximately halved in each model using only one factor. The simpler models, GM1 and GM1MM, which do not include thermodynamics, were not so successful if only a single factor was included (Figure 2A). This is correlated with the statistics in Table 2, where the binding sites modeled by the thermodynamic approach also gave higher χ^2 values than the exact-matching bindings. However, after more factors were added to the models, the picture was more homogeneous (Figure 2B); it seems that simpler models are just as effective in predicting the failure of primers as more complex models. The best model, GM1, helps to achieve a 3-fold (from 17% to 6%) decrease in the failure rate of primers in our dataset.

Description of the major factors

Our models show that one of the major causes of failure of the PCR reaction is an excessive number of primer binding sites. The dynamics of the best factors for failure rate is shown in Figure 3. One can see that alternative binding sites increase the failure rate of the PCR reaction. The effect is very similar for binding sites modeled with or without mismatches and for modeling based on exact matches or thermodynamics (Figure 3A). Allowing a maximum of ten binding sites for both primers in a pair, we can reduce the failure rate from 17% to 10% using only a single factor, which has the greatest impact on the PCR failure rate. To increase it further, we need to include more factors in our models, as described above.

However, the additional factors, such as the GC contents of 3' ends of primers with different word sizes (8 and 16 nt), improve the simpler models to a level similar to the PCR model. Primers with high GC content tend to give higher PCR failure rates (Figure 3B). Although a greater number of predicted PCR products (Figure 3C) increases the PCR failure rate, adding this to the PCR model does not affect the failure rate significantly when the given model is compared with one and four factors (Figure 2). A similar effect was observed with PCR product length.

The GM1 model outperforms RepeatMasker

Primer design can also be conducted from repeat-masked genomic sequences. One purpose of our work was to find a good algorithm for masking regions of

genomic sequences that can lead to the design of failing primers for large genomes. We estimated the failure rate of primers designed to the (1) unmasked genome, (2) RepeatMasker masked genome, (3) GenomeMasker masked genome and (4) genome masked by GM1 with 4 factors, as described in this paper (Table 3). We also estimated the appropriate level of sensitivity for the GM1 model by selecting 1000 random regions from the human genome (each 1000 nucleotides long), masked those regions with each method and tried to design primers for the masked sequences with the Primer3 program. Then we recorded the fraction of masked regions for which at least one primer pair could be designed. This approach helps to select appropriate cutoff levels of parameter values for the GM1 model, with a good balance between successful primer design for any desired region of the genome and the PCR failure rate of the designed primers.

Table 3 shows that using GM1 models with most stringent settings (10% remaining primer candidates) gives a failure rate 2.3 times lower than using RepeatMasker (6% vs. 14%). At these settings GM1 can still design primers into 98.5% of randomly selected 1kb long genomic regions (setup similar to SNP region amplification).

Overall genome masking percentage with RepeatMasker was 50%, with GenomeMasker (max 10 binding sites allowed, masking 1 nucleotide) 52% and with GM1 model (with 4 factors, masking only one nucleotide from 3' end of the repeated word) 81% of nucleotides of human genome. We also selected 1000 random exonic sequences from all known human genes and observed that they are masked even more extensively by GM1 model (86% of nucleotides masked at most stringent settings), but less extensively by RepeatMasker (5% of exon nucleotides masked) or GenomeMasker (28% of exon nucleotides masked). Higher masking of exon regions by our GM1 method may reflect the ability of GM1 to take GC-content of primers into account. Generally GC-rich primers have higher failure rate and therefore GC-rich exon regions are more extensively masked (Figure 3B). If this will cause problems in primer design in exon regions, less stringent settings of GM1 can be used for masking and primer design.

DISCUSSION

The prediction of PCR failure rate on the basis of sequence proves to be an effective way to update current primer design methods. To understand what actually affects the outcome of PCR it is essential to build a model for pre-masking ineffective primer candidates in the genome. With this study, we want to improve the masking methodology instead of methodology of finding PCR products, commonly known as e-PCR. Although our analysis shows that number of PCR products does not influence PCR failure rate significantly, it might still be necessary to predict how many PCR products given primer pair

21

would generate. For this we would suggest using e-PCR programs (GenomeTester, mePCR, isPCR, PRIMEX). A comparison of these programs is shown elsewhere (12).

Our study did not involve statistics of the 3' end base compositions of PCR primers or their combinations with amplicon bases immediately next to the 3' end of the primer (6). This was because the number of experimentally tested primers that we could use in our statistical analysis is small (and so is the number of combinations of base compositions of 3' ends of primers and the sequences of amplicons flanking that 3' end). We very quickly ran into the problem of over-fitting the model with those nucleotide combinations present, and therefore decided to remove them later from our factor list. Nevertheless, we have analyzed separately the effects of the last one and the last two nucleotides of the primer 3' end and the first amplicon nucleotide following primer (Suppl. Figure 1). None of the nucleotide combinations was significant in our datasets although some small individual effects were noticed (Suppl. Table 1). In another study, the most sensitive parameter was the regionalized GC content of the DNA template (7). In our case, this factor (PROD_AUCGC2) had a little significance in training sets, but not in control sets. One possible explanation may be the character of the regions in which the primers are located. Our study includes primers from random regions over the genome (mostly intergenic), whereas Benita's work (7) is based on amplifying human exons, in which the GC content is higher than the average for the human genome.

We have managed to reduce the failure rate of PCR from 17% to 6% with models incorporating four factors. This was achieved with enhanced repeat-masking modeling using any of the GM* models before the primer design process. To improve PCR quality even further, optimization of the reaction protocol and selection or amount of reagents are advised in many studies (2,5,24). Our experiments are based on the human DNA sequence only, but we expect our models to perform with similar efficiency in other large (eukaryotic) genomes.

The primers in this study were not selected randomly. Most of them were designed by Primer3 and some by other primer design programs. The primer design process eliminates many primers with high failure rates (primers with single-nucleotide repeats, extreme GC%, etc.). Our statistical models cannot estimate the influence of factors that had already been filtered out in this process.

The models presented here suggest that using more than one factor can make simpler models as effective as complex ones. The GM1 model contains the maximum number of primer pair binding sites counted with three different word sizes (14, 15 and 16 nt.). Although the most significant factor is MAX16, others complement the model with additional information. This can be compared with the GM1MM model with one factor, which includes binding sites with mismatches. The shorter words in the GM1 model may actually behave like the

mismatch versions of the 16 nt. words and therefore the failure rates are similar to those obtained with more complex models. On the basis of our earlier findings, modeling of primer binding sites by mismatches and exact matches are highly correlated with each other (12). Thus, primers with few exact hits in the genome also have few binding sites with mismatches, and vice versa. Furthermore, the GM1 model can compensate the lack of thermodynamics by evaluating the GC content of the 3' end of the PCR primer. Primers with higher GC contents tend to bind more strongly to alternative sites. Therefore, two primers with the same number of binding sites and different GC contents may have different PCR failure rates. The combined effect of these factors in GM1 creates behavior similar to GM2.

Taking these results together, it is possible to mask genomic DNA regions against possible primer candidates with high failure rates using 100% exact matches only. The current version of our GenomeMasker application supports only single-threaded scanning of genomic DNA, and masking all words with given lengths in the genome that appear over the cutoff value defined by the user. Upgrading of this application should include the "triple-masking" procedure, which means counting words of different sizes simultaneously. The cutoff values for number of binding sites word sizes should be replaced with failure rate probabilities, where the user can define the strictness of the program by choosing desired failure probability cutoff level.

In summary, the number of binding sites is the strongest predictor of PCR failure rate. The GM1 model, based on exact matches with fixed word sizes, is as efficient as more complex models that include thermodynamics and mismatches. Non-unique PCR primers are one major factor that can cause failure of the PCR reaction in experiments with complex genomes. Alternative binding sites create unwanted amplifications, lowering the yield of primers that can hybridize with the desired region or product, and therefore should be avoided whenever possible. This should give us a good opportunity to create better repeat-masking algorithms and reduce the PCR failure rate.

ACKNOWLEDGEMENTS

This work was funded by Enterprise Estonia (EU19730), the Estonian Science Foundation (ETF#6041), target financing grant from Estonian Ministry of Education and Research 0182649s04 and EU FP6 grants "ArraySBS" and "CONCO". The authors would like to thank Asper Biotech Ltd. and Eneli Oitmaa sincerely for conducting the experimental part of this study. The authors also acknowledge Triinu Kõressaar, Priit Palta and Helena Andreson for critical reading of the manuscript.

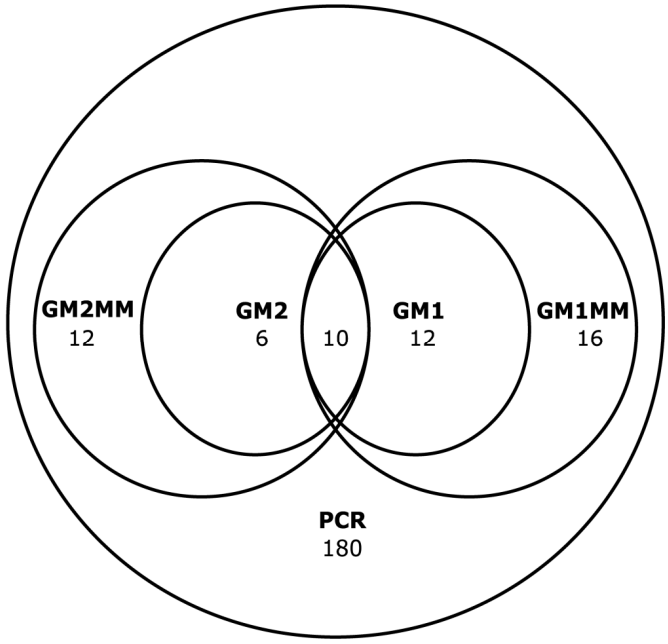
REFERENCES

1. Cobb, B.D. and Clarkson, J.M. (1994) A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Res*, **22,** 3801–3805.
2. Innis, M.A. and Gelfand, D.H. (1990) In Innis, Gelfand, Sninsky and White (eds.), *PCR Protocols*. Academic Press, New York, pp. 3–12.
3. Varadaraj, K. and Skinner, D.M. (1994) Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene*, **140,** 1–5.
4. Haas, S., Vingron, M., Poustka, A. and Wiemann, S. (1998) Primer design for large scale sequencing. *Nucleic Acids Res*, **26,** 3006–3012.
5. Beasley, E.M., Myers, R.M., Cox, D.R. and Lazzeroni, L.C. (1999) In Innis, M. A., D.H., G. and Sninsky, J. J. (eds.), *PCR Applications: Protocols for Functional Genomics*. Academic Press, San Diego, California, USA, pp. 55–72.
6. Yuryev, A., Huang, J., Pohl, M., Patch, R., Watson, F., Bell, P., Donaldson, M., Phillips, M.S. and Boyce-Jacino, M.T. (2002) Predicting the success of primer extension genotyping assays using statistical modeling. *Nucleic Acids Res*, **30,** e131.
7. Benita, Y., Oosting, R.S., Lok, M.C., Wise, M.J. and Humphery-Smith, I. (2003) Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res*, **31,** e99.
8. Yuryev, A., Huang, J., Scott, K.E., Kuebler, J., Donaldson, M., Phillips, M.S., Pohl, M. and Boyce-Jacino, M.T. (2004) Primer design and marker clustering for multiplex SNP-IT primer extension genotyping assay using statistical modeling. *Bioinformatics*, **20,** 3526–3532.
9. Onodera, K. and Melcher, U. (2004) Selection for 3' end triplets for polymerase chain reaction primers. *Mol Cell Probes*, **18,** 369–372.
10. Miura, F., Uematsu, C., Sakaki, Y. and Ito, T. (2005) A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. *Bioinformatics*, **21,** 4363–4370.
11. Chavali, S., Mahajan, A., Tabassum, R., Maiti, S. and Bharadwaj, D. (2005) Oligonucleotide properties determination and primer designing: a critical examination of predictions. *Bioinformatics*, **21,** 3918–3925.
12. Andreson, R., Reppo, E., Kaplinski, L. and Remm, M. (2006) GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics*, **7,** 172.
13. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **35,** D5–12.
14. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, **32,** D493–496.
15. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25,** 3389–3402.
16. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, **16,** 418–420.

17. SantaLucia, J., Jr. and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct*, **33,** 415–440.
18. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31,** 3406–3415.
19. Kaplinski, L., Andreson, R., Puurand, T. and Remm, M. (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, **21,** 1701–1702.
20. Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418,** 544–548.
21. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, **132,** 365–386.
22. Miller, S.A., Dykes, D.D. and Polesky, H.F. (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*, **16,** 1215.
23. Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*, **135,** 370–384.
24. SantaLucia, J., Jr. (2007) Physical principles and visual-OMP software for optimal PCR design. *Methods Mol Biol*, **402,** 3–34.

22

FIGURES



Total: PCR=236 **GM2MM**=28 **GM2**=16 **GM1MM**=38 **GM1**=22

Figure 1. The distribution of factors between different model types.

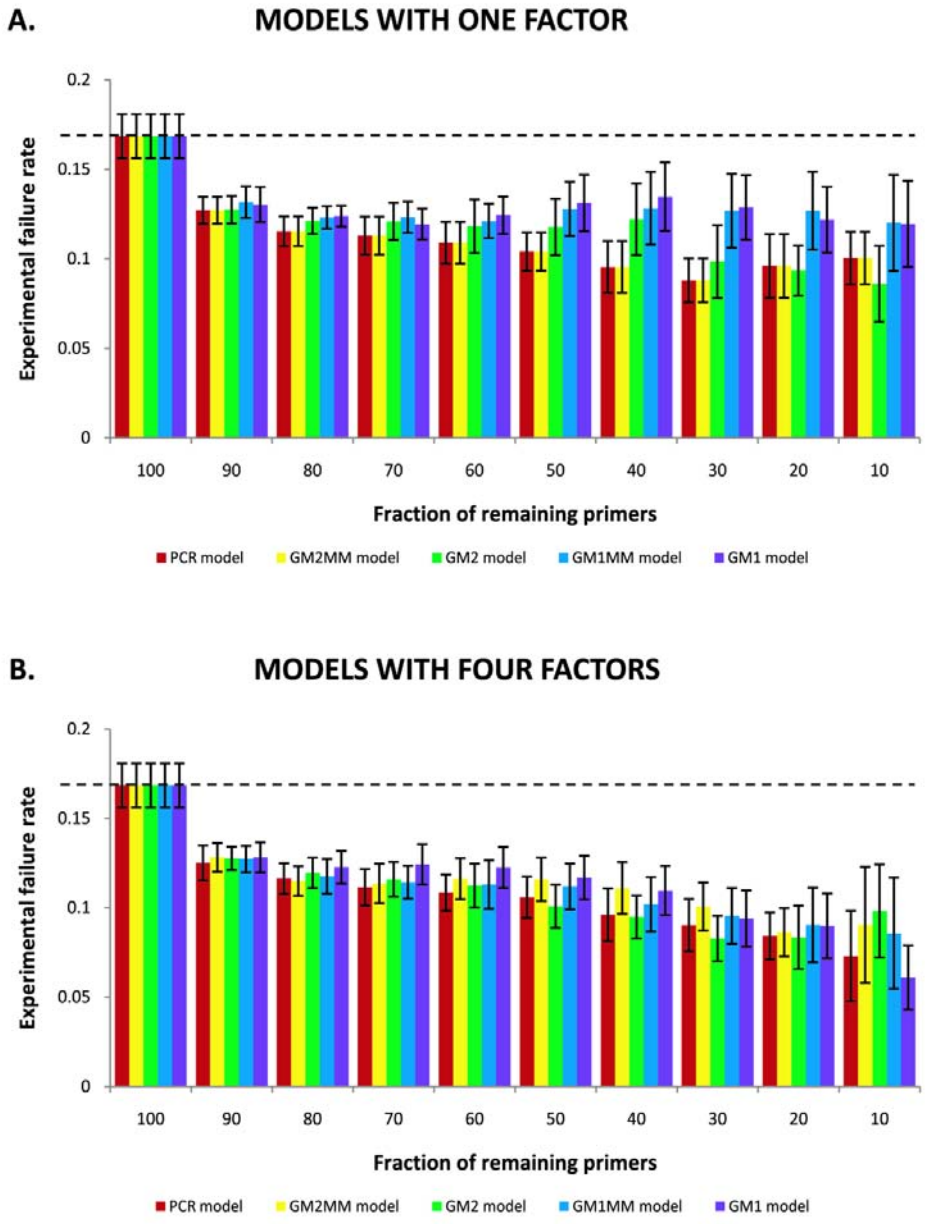


Figure 2. Comparison of five model types at different sensitivity levels. The figure illustrates the efficiency of the models in predicting the actual failure rate at several cutoff levels with a single factor (A) or four factors (B) included in the model. The average failure rate of the remaining primer pairs in 10 control sets using the given cutoff. Dashed line defines the average PCR failure rate of the whole dataset before applying any model. Error bars show 95% confidence limits for the real mean.

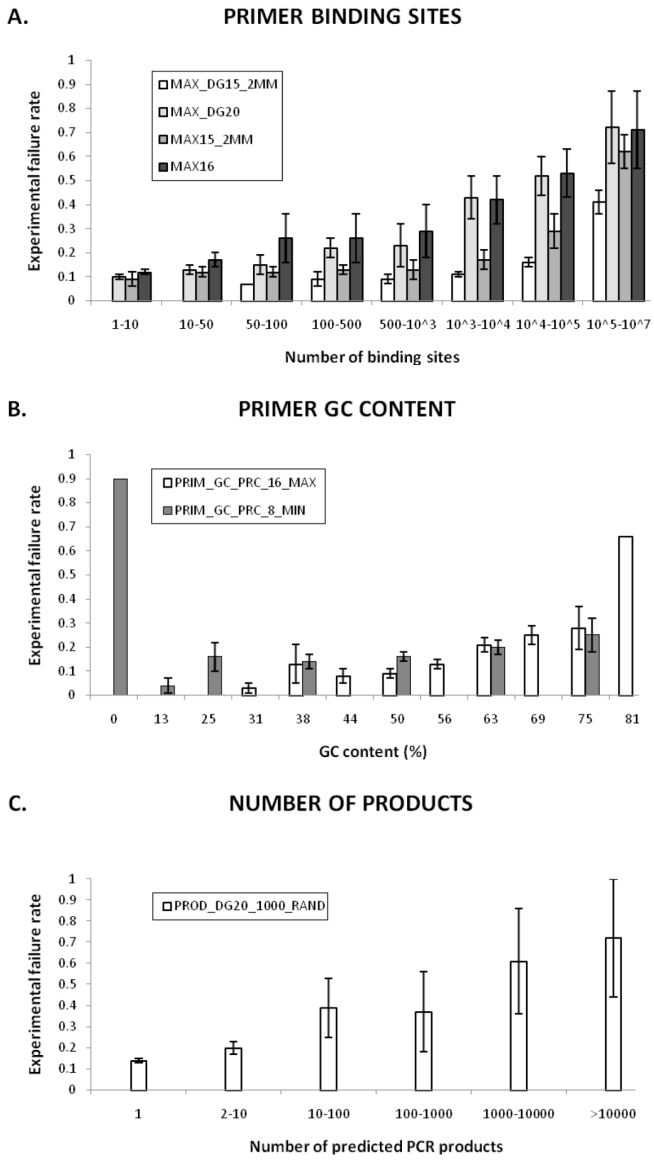


Figure 3. The relationship between some of the statistically significant factors and the PCR failure rate. The binding sites of the primer pairs were counted with the most descriptive factors for each model: (A) maximum number of hits with two mismatches allowed and free energy level ≤ –15 kcal/mol (PCR and GM2MM models), no mismatches allowed and free energy level ≤ –20 kcal/mol (GM2 model), two mismatches allowed and word size = 15 nt. (GM1MM model), no mismatches allowed and word size = 16 nt. (GM1 model). The effects of the primer GC content (B) and the number of predicted PCR products (C) are also shown. Error bars show 95% confidence limits for the real mean.

TABLES

Table 1. The complete list of factors used in study for building models.

Factor description	Factor name	GM1	GM1MM	GM2	GM2MM	PCR	Number of factors
The number of binding sites of PCR primers (exact, with one and two mismatches allowed) with different word sizes from the 3' end and from random positions in the primer sequence.	MAX/MIN[8,10,12,14,15,16];	+	+			+	12
	MAX/MIN _FULL;					+	2
	MAX/MIN[8,10,12,14,15,16]_RAND;					+	12
	MAX/MIN[12,14,15,16]_1MM;		+			+	8
	MAX/MIN_FULL_1MM;					+	2
	MAX/MIN[12,14,15,16] _1MM _RAND;		+			+	8
	MAX/MIN[12,14,15,16]_2MM;					+	8
	MAX/MIN_FULL_2MM;					+	2
	MAX/MIN[12,14,15,16] _2MM _RAND					+	8
The number of binding sites of PCR primers (exact, with one and two mismatches allowed) with variable word sizes from the 3' end and from random positions in the primer sequence. The word size for each primer is extended until three different free energy levels are achieved: ΔG < -10, -15, -20 kcal/mol.	MAX/MIN_DG[10,15,20];			+	+	+	6
	MAX/MIN_DG[10,15,20]_RAND;					+	6
	MAX/MIN_DG[10,15,20]_1MM;				+	+	6
	MAX/MIN_DG[10,15,20]_1MM RAND;					+	6
	MAX/MIN_DG[10,15,20]_2MM;				+	+	6
	MAX/MIN_DG[10,15,20]_2MM _RAND					+	6
The number of all binding sites of PCR primers (exact, with one and two mismatches allowed) counted with NCBI BLASTN (-F F).	[MAX,MIN]_BLASTALL					+	2
PCR primer length	PRIM_LENGTH_[MAX,MIN]					+	2
GC content of PCR primer with different word sizes from the 3' end and full primer	PRIM_GC_PRC_[8,12,16]_[MAX,MIN];	+	+	+	+	+	6
	PRIM_GC_PRC_[MAX,MIN]					+	2
The free energies of different subsequences from the primer 3' end	PRIM_ DG[3,4,5,6,7,8,9]_[MAX,MIN]	+	+	+	+	+	14
DUST score of PCR primer	PRIM_DUS_[MAX,MIN]					+	2
The strongest free energies of the dimers of primers alone and in pairs using local and global alignment approaches	MAX/MIN_PRIM_END1;					+	2
	PRIM_PAIR_END1;					+	1
	MAX/MIN_PRIM_END2;					+	2
	PRIM_PAIR_END2;					+	1
	MAX/MIN_PRIM_ANY;					+	2
	PRIM_PAIR_ANY					+	1

23

Factor description	Factor name	GM1	GM1MM	GM2	GM2MM	PCR	Number of factors
The strongest secondary structure of the PCR primers in a given pair predicted with MFOLD at 55^{o} C	[MAX,MIN]_PRIM_MFOLD					+	2
The melting temperature of the primer, difference of melting temperatures between the two primers in a given pair and the difference between annealing (used in PCR experiments) and melting temperature	TM_[MAX,MIN]; TM_DIFF; TM_TA_[MAX,MIN]_DIFF					+ + +	2 1 2
Total number of SNPs in both primers and the position of the SNP closest to the 3' end	NO_OF_SNPS; ALL_POS_FROM_3_END; NO_OF_VALID_SNPS; VALID_POS_FROM_3_END	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	1 1 1 1
The terminal and last two nucleotides of primer sequence, also the first nucleotide of amplicon following the primer sequence. These are categorical values (0 – given nuc. is not present in both primers, 1 – is present at least in one primer, 2 – is present in both primers).	PRIM_LAST_ONE_NUC_[A,C,G,T]; PRIM_LAST_TWO_NUC_[AA,AC,AG,AT,CC,CG,CT,GG,GT,TT]; PROD_FIRST_ONE_NUC_[A,C,G,T]					+ + +	4 10 4
The number of predicted products with maximum length of 1000, 3000 and 10000 nt for exact binding sites with different word sizes from the 3' end and from random positions in the primer sequence.	PROD[8,10,12,14,15,16]_1000; PROD_FULL_1000; PROD[8,10,12,14,15,16]_1000_RAND; PROD[8,10,12,14,15,16]_3000; PROD_FULL_3000; PROD[8,10,12,14,15,16]_3000_RAND; PROD[8,10,12,14,15,16]_10000; PROD_FULL_10000; PROD[8,10,12,14,15,16]_10000_RAND					+ + + + + + + + +	6 1 6 6 1 6 6 1 6
The number of predicted products with maximum length of 1000, 3000 and 10000 nt for exact binding sites with variable word sizes from the 3' end and from random positions in the primer sequence. The word size for each primer is extended until three different free energy levels are achieved: ΔG < -10, -15, -20 kcal/mol.	PROD_DG[10,15,20]_1000; PROD_DG[10,15,20]_1000_RAND; PROD_DG[10,15,20]_3000; PROD_DG[10,15,20]_3000_RAND; PROD_DG[10,15,20]_10000; PROD_DG[10,15,20]_10000_RAND					+ + + + + +	3 3 3 3 3 3

Factor description	Factor name	GM1	GM1MM	GM2	GM2MM	PCR	Number of factors
PCR product length	PROD_LENGTH					+	1
GC content of PCR product	PROD_GC_PRC					+	1
Area under the GC curve and above 65% of the PCR product (Benita *et al.*, 2003)	PROD_AUCGC					+	1
Number of GC windows with values above 65% divided by the length of the PCR product (x100) (Benita *et al.*, 2003)	PROD_RATIOGC_100					+	1
PROD_AUCGC x PROD_RATIOGC (Benita *et al.*, 2003)	PROD_AUCGC2					+	1
The strongest secondary structure of PCR product predicted with MFOLD at 55^{o} C	PROD_MFOLD_55					+	1
Percentage of masked nucleotides of PCR product using DUST	PROD_DUST_PRC					+	1
Percentage of masked nucleotides of PCR product using RepeatMasker with different sensitivity parameters (-s, -q, -qq)	PROD_RMs_PRC; PROD_RMq_PRC; PROD_RMqq_PRC					+ + +	1 1 1
Percentage of masked nucleotides of PCR product using GenomeMasker with different word sizes (exact matches)	PROD_GM[8,10,12,14,16]_PRC					+	5
Total number of factors:							**236**

Factors marked by '+' under a model are used in the building of this model.

Table 2. List of the best factors (top 4) and the corresponding one-degrees-of-freedom chi-squares ($\chi^2(1)$) from the GENMOD Type I analysis using whole dataset. All factors are significant at $p<0.0001$.

Factor name	χ^2 (1)	Model
MAX_DG15_2MM	4862	PCR
MAX_DG15_RAND*MAX_DG15_RAND	1374	
PROD_DG20_1000_RAND*PROD_DG20_1000_RAND	378	
PROD_LENGTH*PROD_LENGTH	298	
MAX_DG15_2MM	4862	GM2MM
MAX_DG15_1MM*MAX_DG15_1MM	1091	
MAX_DG20_2MM	244	
MAX_DG20_1MM*MAX_DG20_1MM	262	
MAX_DG20	4085	GM2
MAX_DG15*MAX_DG15	1106	
PRIM_GC_PRC_8_MIN	386	
MIN_DG20*MIN_DG20	277	
MAX15_2MM	2854	GM1MM
MAX12_1MM*MAX12_1MM	1681	
MAX12	1291	
PRIM_GC_PRC_16_MAX	789	
MAX16	2507	GM1
MAX15*MAX15	2394	
PRIM_GC_PRC_16_MAX	1126	
MAX14	272	

Asterisks in factor names mark the polynomial regression of given independent variable. χ^2 values illustrate the estimated simultaneous (Type I) effects of the best four factors on each model.

Table 3. Comparison of our GM1 model with other masking methods.

	No masking	Repeat-Masker	Genome-Masker	GM1 with 4 factors				
				10%	20%	30%	40%	50%
Failure rate[a]	16.9	13.8	10.2	6.0	8.9	9.5	11.2	11.4
Primer design possible[b]	100	69.3	96.4	98.5	99.3	99.6	99.6	99.7
Sequence masked[c]								
in genome		49.5	52.2	80.7	59.7	51.4	45.5	35.6
in introns	0	12.5	38.8	83.7	63.4	54.5	47.7	37.6
in exons		4.5	28.2	86.0	69.7	61.1	54.0	40.4

[a] fractions of failing primer pairs after using given masking method calculated from the experimental data of 1314 primer pairs.

[b] fraction of masked genomic regions for which at least one primer pair could be designed using 1000 random regions from the human genome (each 1000 nt long). Primer3 was used to design primer pairs.

[c] fraction of masked nucleotides from three different random sequence sets: 1000 genomic regions (each 1000 nt long), 1000 exonic sequences (average length 150 nt) and 1000 intronic sequences (average length 400 nt).

With GenomeMasker and GM1 we have used an option to mask only one nucleotide from 3' end of the repeated word.

CURRICULUM VITAE

Reidar Andreson

Date and place of birth: 7. May 1978, Tartu, Estonia
Address: Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu
Riia str. 23, 51010, Tartu, Estonia
Phone: +372 7375002
E-mail: reidar.andreson@ut.ee

Education and professional employment

1985–1996	Ülenurme Gymnasium
1996–2000	B.Sc University of Tartu, Institute of Molecular and Cell Biology
2000–2002	M.Sc University of Tartu, Institute of Molecular and Cell Biology
2000	European Bioinformatics Institute (EBI), Hinxton, UK, guest student
2000–2002	Asper Biotech Ltd., IT specialist
2002–	Ph.D student in Department of Bioinformatics, University of Tartu, Institute of Molecular and Cell Biology
2002–2004	BioData Ltd., IT specialist
2004–2008	University of Tartu, research scientist
2008–	Estonian Biocentre, research scientist

Scientific work

My research projects are associated with development of specific bioinformatics methods and algorithms for genotyping applications. These include improvements in DNA identity search and repeat masking algorithms. Later on, I have been modeling the prediction of PCR failure rate, identifying the most important sequence-based factors influencing PCR and trying to implement these models into current masking algorithms.

List of publications

1. **Andreson, R**, Möls, T, Remm, M. (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Research.* (accepted)
2. **Andreson, R**, Kaplinski, L, Remm, M. (2007) Fast masking of repeated primer binding sites in eukaryotic genomes. *Methods Mol Biol.* 2007;402:201–18. Review.
3. **Andreson, R**, Puurand, T, Remm, M. (2006) SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. *Nucleic Acids Research*, 34, W651–W656.
4. **Andreson, R**, Reppo, E, Kaplinski, L, Remm, M (2006) GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics*, 7, 172.
5. Kaplinski, L, **Andreson, R**, Puurand, T, Remm, M. (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, 21(8), 1701–1702.

ELULOOKIRJELDUS

Reidar Andreson

Sünniaeg ja koht:	7. mai 1978, Tartu, Eesti
Aadress:	Molekulaar ja Rakubioloogia Instituut, Tartu Ülikool Riia mnt. 23, 51010, Tartu, Eesti
Telefon:	+372 7375002
E-mail:	reidar.andreson@ut.ee

Haridus ja erialane teenistuskäik

1985–1996	Ülenurme Gümnaasium
1996–2000	Tartu Ülikooli Bioloogia-Geograafia teaduskond, biotehnoloogia ja biomeditsiini eriala, B.Sc
2000–2002	Tartu Ülikooli Bioloogia-Geograafia teaduskond, bioinformaatika eriala, M.Sc
2000	Euroopa Bioinformaatika Instituut (EBI), Hinxton, UK, külalisüliõpilane
2000–2002	AS Asper Biotech (IT spetsialist)
2002–	Tartu Ülikooli Loodus- ja tehnoloogiateaduskond teaduskond, bioinformaatika eriala, doktorant
2002–2004	BioData OÜ (IT spetsialist)
2004–2008	Tartu Ülikool, teadur
2008–	Eesti Biokeskus, teadur

Teadustegevus

Teadustöö on seotud bioinformaatika alaselt genotüpiseerimiseks vajalike algoritmide ja meedodite arendamisega. Algselt tegelesin DNA järjestuste sarnasuse ja identsuse otsingu meetodite uurimisega. Hiljem lisandus sellele PCR edukuse hindamine ja spetsiaalsete mudelite loomine eesmärgiga arendada välja efektiivsemad korduste maskeerimise algoritmid.

Publikatsioonid

1. **Andreson, R**, Möls, T, Remm, M. (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Research.* (accepted)
2. **Andreson, R**, Kaplinski, L, Remm, M. (2007) Fast masking of repeated primer binding sites in eukaryotic genomes. *Methods Mol Biol.* 2007;402:201–18. Review.
3. **Andreson, R**, Puurand, T, Remm, M. (2006) SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. *Nucleic Acids Research*, 34, W651–W656.
4. **Andreson, R**, Reppo, E, Kaplinski, L, Remm, M (2006) GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics*, 7, 172.
5. Kaplinski, L, **Andreson, R**, Puurand, T, Remm, M. (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, 21(8), 1701–1702.

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro.* Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous crassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan.** Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.

41. **Sulev Ingerpuu.** Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.
42. **Veljo Kisand.** Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa.** Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa.** Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik.** Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo.** Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo.** Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots.** Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero.** Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees.** Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks.** Cholecystokinin (CCK) — induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and erotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna.** Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro.** Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane.** Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm.** Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg.** Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild.** The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu.** Studies of the TOL plasmid transcription factor XylS. Tartu 2000. 88 p.

26

61. **Dina Lepik.** Modulation of viral DNA replication by tumor suppressor protein p53. Tartu 2000. 106 p.
62. **Kai Vellak.** Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu 2000. 122 p.
63. **Jonne Kotta.** Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu 2000. 160 p.
64. **Georg Martin.** Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000. 139 p.
65. **Silvia Sepp.** Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira.** On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000. 96 p.
67. **Priit Zingel.** The role of planktonic ciliates in lake ecosystems. Tartu 2001. 111 p.
68. **Tiit Teder.** Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu 2001. 122 p.
69. **Hannes Kollist.** Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu 2001. 80 p.
70. **Reet Marits.** Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu 2001. 112 p.
71. **Vallo Tilgar.** Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002. 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002. 108 p.
73. **Liina Eek-Piirsoo.** The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002. 74 p.
74. **Krõõt Aasamaa.** Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002. 110 p.
75. **Nele Ingerpuu.** Bryophyte diversity and vascular plants. Tartu, 2002. 112 p.
76. **Neeme Tõnisson.** Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002. 124 p.
77. **Margus Pensa.** Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003. 110 p.
78. **Asko Lõhmus.** Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003. 168 p.
79. **Viljar Jaks.** p53 — a switch in cellular circuit. Tartu, 2003. 160 p.
80. **Jaana Männik.** Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003. 140 p.
81. **Marek Sammul.** Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003. 159 p

82. **Ivar Ilves.** Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003. 89 p.
83. **Andres Männik.** Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003. 109 p.
84. **Ivika Ostonen.** Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003. 158 p.
85. **Gudrun Veldre.** Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003. 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004. 159 p.
87. **Aare Abroi.** The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004. 135 p.
88. **Tiina Kahre.** Cystic fibrosis in Estonia. Tartu, 2004. 116 p.
89. **Helen Orav-Kotta.** Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004. 117 p.
90. **Maarja Öpik.** Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004. 175 p.
91. **Kadri Tali.** Species structure of *Neotinea ustulata*. Tartu, 2004. 109 p.
92. **Kristiina Tambets.** Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004. 163 p.
93. **Arvi Jõers.** Regulation of p53-dependent transcription. Tartu, 2004. 103 p.
94. **Lilian Kadaja.** Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004. 103 p.
95. **Jaak Truu.** Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004. 128 p.
96. **Maire Peters.** Natural horizontal transfer of the *pheBA* operon. Tartu, 2004. 105 p.
97. **Ülo Maiväli.** Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004. 130 p.
98. **Merit Otsus.** Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004. 103 p.
99. **Mikk Heidemaa.** Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004. 167 p.
100. **Ilmar Tõnno.** The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004. 111 p.
101. **Lauri Saks.** Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004. 144 p.

102. **Siiri Rootsi.** Human Y-chromosomal variation in European populations. Tartu, 2004. 142 p.
103. **Eve Vedler.** Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover.** Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 126 p.
105. **Helen Udras.** Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005. 100 p.
106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005. 162 p.
107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005. 162 p.
108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005. 156 p.
109. **Kaasik, Krista.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005. 121 p.
110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005. 112 p.
111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005. 103 p.
112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005. 112 p.
113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 106 p.
114. **Mait Metspalu.** Through the course of prehistory in india: tracing the mtDNA trail. Tartu, 2005. 138 p.
115. **Elin Lõhmussaar.** The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006. 124 p.
116. **Priit Kupper.** Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006. 126 p.
117. **Heili Ilves.** Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006. 120 p.
118. **Silja Kuusk.** Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006. 126 p.
119. **Kersti Püssa.** Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006. 90 p.
120. **Lea Tummeleht.** Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006. 94 p.
121. **Toomas Esperk.** Larval instar as a key element of insect growth schedules. Tartu, 2006. 186 p.

122. **Harri Valdmann.** Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers.** Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli.** Gata3 and Gata2 in inner ear development. Tartu, 2007. 123 p.
125. **Kai Rünk.** Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007. 143 p.
126. **Aveliina Helm.** Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007. 89 p.
127. **Leho Tedersoo.** Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007. 233 p.
128. **Marko Mägi.** The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007. 135 p.
129. **Valeria Lulla.** Replication strategies and applications of Semliki Forest virus. Tartu, 2007. 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007. 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007. 171 p.
132. **Tatjana Krama.** Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007.
133. **Signe Saumaa.** The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007. 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007. 96 p.
135. **Priit Kilgas.** Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007. 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007. 95 p.
137. **Kärt Padari.** Protein transduction mechanisms of transportans. Tartu, 2008. 128 p.
138. **Siiri-Lii Sandre.** Selective forces on larval colouration in a moth. Tartu, 2008. 125 p.
139. **Ülle Jõgar.** Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008. 99 p.
140. **Lauri Laanisto.** Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008. 133 p.