DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
216

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
216

TRIINU KÕRESSAAR

Improvement of PCR primer design for detection of prokaryotic species

TARTU UNIVERSITY PRESS

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

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in molecular diagnostics) on 1.06.2012 by the Council of the Institute of Molecular and Cell Biology, University of Tartu.

Supervisor: Prof. Maido Remm
Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Estonia

Opponent: Martine Petronella Bos

Commencement: Room No 217, 23 Riia Str., Tartu, on August 23th 2012, at 10.00

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

ISSN 1024–6479
ISBN 978–9949–32–042–4 (trükis)
ISBN 978–9949–32–043–1 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 330

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The current dissertation is based on the following papers referred to by Roman numbers:

I. Koressaar T, Jõers K and Remm M (2007). Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10): 1289–91

II. Koressaar T and Remm M (2009). Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms. Bioinformatics 25(11): 1349–55

III. Koressaar T and Remm M (2012). Characterization of species-specific repeats in 613 prokaryotic species. DNAResearch, published online February 24, 2012

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

Ref I found the most precise method for calculating the melting temperature of an oligo, renewed the code, and participated in the writing of manuscript.

Ref II participated in developing the method, implemented the method in programming language, designed PCR primers, conducted statistical analyses, implemented the web client and participated in the writing of manuscript.

Ref III found repetitive sequences with the method published in Ref II, conducted the analysis of species-specific repeats, and participated in the writing of manuscript.

LIST OF ABBREVIATIONS

ANI	Average Nucleotide Identity
BIME	Bacterial Interspersed Mosaic Elements
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DDH	DNA-DNA Hybridization
dNTP	deoxyribonucleoside-triphosphates
ERIC	Enterobacterial Repetitive Intergenic Consensus Elements
ICE	Integrative Conjugative Elements
indel	insertion or deletion
IRU	Intergenic Repeat Units
IS	Insertion Sequence
MGE	Mobile Genetic Element
MITE	Miniature Inverted Transposable Element
NN	Nearest-Neighbour
PCR	Polymerase Chain Reaction
PU	Palindromic Units
qPCR	quantitative PCR
REP	Repetitive Extragenic Palindromic sequences
T_m	melting Temperature

INTRODUCTION

Despite of widespread use of DNA sequencing, PCR technology is still widely applied in many practical fields. In many cases, new PCR primers identifying pathogenic microbes are designed instead of existing PCR primers that lack sensitivity or specificity. Besides, primers for identifying some new prokaryotic species are frequently required. Thus constant improvement of PCR primer design is essential. One should have in mind that PCR primer design is not exclusively a generation of PCR primer sequences but rather a process comprising several steps (choosing a target sequence, examination of parameter values applied to design primer sequences, testing secondary structures of primer/product sequences, testing specificity of primers). Advances in PCR primer design introduced here comprise first, enhancements of PCR primer design program Primer3, second, a new approach to prokaryotic PCR primer design and third, the study of variable characteristics of microbial genome repetitive sequences to provide alternative targets for primer design. The re-examination of the PCR primer design algorithms and programs at times is important. For example, until 2007 widely used primer design program Primer3 was using primer melting temperature formulae, which included some specific constants from papers published in 1965–1990 (Koressaar and Remm, 2007). Subsequently, several improved, more accurate formulae and constants were suggested. Furthermore, the rising number of complete prokaryotic genome sequences allows delevopment of automatic solutions for primer design. Alternative target sequences for prokaryotic primer design are necessary, as many current PCR primers lack good sensitivity (e.g. the amount of DNA in clinical samples is limited) or in the case of, e.g. 16S rDNA target sequences, which lack specificity (sequences are too conserved between closely related species). Here we propose two methods, one for finding prokaryotic species-specific repetitive sequences, and the other for designing prokaryotic species-specific PCR primers on the repeats. Finally, we have exhaustively analysed prokaryotic species-specific repeats to determine their main characteristics, their diversity and their possible roles. This analysis gives some assurance when using the species-specific sequences in primer design and also provides many other interesting issues concerned to species-specific repeats.

In the literature section of this thesis, the first part is devoted to PCR and primer design. An overview of the most important parameters that influence PCR outcome is given. Description of primer design program Primer3 with the principles of calculating the melting temperature of short oligos is also provided. Furthermore, prokaryotic sequences used in primer design are introduced, with a short discussion on them and the list of *ab initio* programs for finding repetitive sequences. Finally, common prokaryotic repetitive sequences have been characterized.

In the research part of this thesis, improvements incorporated into primer design program Primer3 are first introduced. These improve the accuracy of predicting the melting temperature of primers. Next, the methods for finding

species-specific repetitive sequences and species-specific PCR primers are addressed. While the first method includes the BLAST similarity search algorithm, the second includes Primer3. This part also includes a description of the experimental analysis, showing that the increase in the number of copies of target sequences in the genome raises the sensitivity of PCR. Finally, an overview of a comprehensive study of species-specific repetitive sequences is provided. This covers the characterization of species-specific repeats and the classification of repeats by functional analysis.

1. REVIEW OF LITERATURE

1.1. Polymerase chain reaction

1.1.1. PCR and parameters influencing the PCR process

The polymerase chain reaction (PCR) is a valuable tool in molecular biology that is used in research and in routine applications. The method was developed by Mullis on 1983. Polymerase chain reaction amplifies a specific DNA sequence very quickly (Mullis et al., 1986). It is widely used because it is sensitive, robust, fast, simple to use and easy to optimize. Besides the scientific labs, PCR is also currently applied in clinical, veterinary, food diagnostics, and environmental monitoring.

There have been developed several different PCR techniques, many of them being used routinely in diagnostics, for example multiplex-PCR and real-time PCR. Multiplex PCR is similar to conventional PCR but uses more than one primer pair in each reaction tube. Thus it simultaneously allows identification of many target sequences. Because of its nature, cost-effectiveness and speed, multiplex-PCR is frequently applied in the diagnostics of pathogens (Chamberlain et al., 1988). Today, conventional PCR methods are mostly replaced by real-time PCR (O'Connor and Glynn, 2010). Real-time PCR is more sensitive than conventional due to quantification of accumulating PCR product in real-time (Heid et al., 1996). For detecting single-stranded PCR product, different fluorescently labelled sequence-specific probes are available, such as TaqMan™ probes and HybProbes. For instance, TaqMan™ technology employs probe that has a fluorescent label attached to the 5′ end and a quencher at the 3′ end. During the PCR reaction, the probe is hybridized to the single stranded PCR product. DNA polymerase cleaves the probe separating the fluorescent label from the quencher and causing the emitting of signal (O'Connor and Glynn, 2010). As it is more beneficial to identify several pathogens in the same reaction tube, some qPCR (quantitative PCR) technologies have adapted to multiplexing. For instance, using TaqMan™ chemistry, several sequence-specific probes can be labelled with different fluorophores, and different targets can be amplified and quantified within a single reaction (Smith and Osborn, 2009; Wong and Medrano, 2005).

Anyone designing a PCR primer pair must keep in mind that PCR primers must meet several criteria. With a very careful/detailed computational primer design, minimal PCR optimization is required in laboratory. In general, there are mainly three circumstances to consider in design of PCR primer sequences – (I) prediction of accurate melting temperature of a primer-target duplex, (II) correct prediction of the secondary structures of both, primer sequences (dimers and hairpins) and target sequence and (III) correct prediction of secondary binding sites, where PCR primers could anneal (discussed further). Many parameters affect the primer melting temperature, the formation of both, primer dimers and primer hairpins, and the annealing of primer to secondary binding

sites. Although there are many other circumstances that affect the success of PCR (the length of PCR primers and the length of PCR product, the base composition, GC-content and thermodynamic stability of primer 3' end), these cannot currently be included in primer design because of a lack of full understanding the mechanism behind them (Haas et al., 1998; Kwok et al., 1990; Li et al., 1997; Miura et al., 2005; Onodera and Melcher, 2004). Some target sequences enable to design of PCR primers, which result a successive PCR. However, it is not always possible to choose between different target sequences and thus the sequence required to amplify may flank with low-complexity sequences. In this case, greater effort is needed in primer design (Kampke et al., 2001).

(I) Melting temperature of an oligonucleotide relies on a three major factors: the concentration of oligonucleotide, the concentration of salt cations, and the base composition of oligonucleotide sequence. High hybridizing DNA concentrations favour duplex formation and increase the melting temperature. Monovalent and divalent cations have stabilizing effects on DNA duplex formation, and therefore increase the melting temperature. It is a common understanding that different cations may have different effect on melting temperature. However, various views exist on the influence of the same cations on melting temperature (Nakano et al., 1999; Cheng et al., 2006; Lyubartsev and Laaksonen, 1998; Owczarzy et al., 2008; Owczarzy et al., 2004; SantaLucia, 1998; Schildkraut and Lifson, 1965). The base composition of oligonucleotide determines the nearest-neighbour interactions (free energy values) considered for calculating melting temperature (Marky and Breslauer, 1982).

(II/III) In general, two types of primer secondary structures exist – primer-primer dimers and primer monomers, known as primer hairpins. The occurrence of primer self-annealing (primer-primer dimers) can result in short unexpected PCR products, while the occurrence of primer self-end-annealing (hairpins) can result in a low yield of PCR product (Chou et al., 1992; Singh et al., 2000). As primers can hybridize to sequence without full complementarity, the correct prediction of primer secondary binding sites is important. Primers annealing to secondary sites can cause amplification of false products, and can also lower the yield of correct PCR product (Johnson, 2000).

Two possibilities are mainly used to predict secondary structures of sequences (DNA duplexes and hairpins). One is simple calculation of the Watson-Crick matches between two hybridizing DNA sequences and therefore finding a score for the predicted structure. The second is based on prediction of the transition melting temperature of DNAs potential secondary structures. The latter is based on the nearest-neighbour model, which helps to find energetically stable structures at given conditions (SantaLucia et al., 1996). The nearest-neighbour model uses 2-state approximation (no intermediate states are assumed). At given conditions the most probable secondary structure is with the smallest free energy value. Many computer programs exist that calculate hairpin or duplex formation based on the nearest-neighbour model. For example, UNAFold predicts folding and hybridization of two strands of DNA or RNA sequence (Markham and Zuker, 2008). Another program, FastaGrep, also finds

all possible binding sites of an oligo sequenceby thermodynamic alignment (Kaplinski L, unpublished, http://bioinfo.ut.ee/download/).

1.1.2. Overview of Primer3

Primer design program Primer3 is an expansion of primer design program Primer 0.5 written by Steve Lincoln, Mark Daly and Eric S. Lander. Primer3 was originally written by Helen J. Skaletsky and Steve Rozen in 1991. Today, Primer3 is an open software development project hosted on SourceForge, and Primer3 is published under the GNU General Public Licence by the Free Software Foundation (http://www.gnu.org/licenses/gpl-2.0.txt). There are several developers of Primer3 from different places of the world, including the original author, Steve Rozen, and the author of this work, Triinu Kõressaar (http://sourceforge.net/projects/primer3/). While Primer3 is freeware and can be used over the web as well as included into automatic primer design, it is widely used. It is flexible and fast; for example, it is platform-independent and it can be used through command line in large scale projects. Primer3 also enables the design of hybridization oligos. Because of all these characteristics, many specific primer design programs apply Primer3 (Gadberry et al., 2005; Andreson et al., 2006; Srivastava et al., 2008).

Primer3 has a wide range of user adjustable parameters that all have optimal default values (Rozen and Skaletsky, 2000), e.g. the maximum, minimum and optimum values for each of followings: melting temperature of primer, GC% of primers and length of primers. Also, the user can choose a formula that is applied for calculating melting temperature (next to the salt correction formula that is used for calculating the melting temperature) of primers, can adjust values for concentration of monovalent cations, divalent cations and hybridizing DNA sequences, can adjust the cut-off values for primers/oligos self-annealing and for primers self-end-annealing, and the cut-off value for primers/oligos hairpins and primers mispriming to template. The method that is used for calculating primer/oligo dimers and their hairpins (either method- that is based on thermodynamic approach or method- that is based on counting Watson-Crick pairings), has to be also chosen by user.

In the primer design process, Primer3 calculates a penalty for every primer pair. The pair with the smallest penalty is the best fit with user-determined parameter values. For calculating a penalty, Primer3 sums sub-penalties which, in turn, are calculated for every parameter by a specific formula that uses user-determined weights and constraints (user-determined parameter values) for a certain parameter.

1.1.3. Methods for calculating the melting temperature of an oligo

Two-state melting temperature model is assumed for short oligonucleotide sequences. According to this model, fully intact duplex and completely dissociated single strands are the only states that are populated by every strand throughout the entire melting transition. Thus, the melting temperature (T_m) is the temperature at which 50% of the molecules are single-stranded and 50% are double-stranded. This means that partially melted duplexes are present in negligible amounts (Owczarzy et al., 1997).

Several methodologies exist for calculating DNA/DNA homoduplexes (up to 60nt) melting temperature. The first widely used and the most trivial method is so called *basic* or the Wallace and Ikatura rule, which considers only G+C content of hybridizing oligos (Wallace et al., 1979).

$$T_m\,(°C) = 2 * (A+T) + 4 * (G+C) \tag{1}$$

More precise formulae use the nearest-neighbour approach for calculating the melting temperature of an oligo. The nearest-neighbour model for nucleic-acids assumes that the stability of a given base-pair depends on the identity and the orientation of neighbouring base-pair. Also nearest-neighbour models consider the concentration of monovalent and divalent cations, and the concentration of hybridizing oligonucleotides (SantaLucia, 1998).

Many tables with nearest-neighbour parameter values and formulas for both (oligo melting temperature calculation and salt correction) have been published (Freier et al., 1986; Breslauer et al., 1986; Sugimoto et al., 1995; Sugimoto et al., 1996; SantaLucia et al., 1996; SantaLucia, 1998; SantaLucia and Hicks, 2004; Xia et al., 1998; Owczarzy et al., 2004). There have been made several comparative studies to find out the most precise formula for calculating the melting temperature of a duplex (Chavali et al., 2005; Panjkovich and Melo, 2005; Owczarzy et al., 2004). However, the general formula for duplex melting temperature calculation is:

$$T_m = \frac{\sum\left(\Delta H_d^o\right)}{\sum\left(\Delta S_d^o\right) + R * \ln\left(C_T / x\right)} \tag{2}$$

where ΔH_d^o and ΔS_d^o are enthalpy and entropy changes respectively, R is the gas constant (1.987cal/K*mol), C_T is the total molar strand concentration and x equals 4 or 1 for nonself-complementary duplexes and for self-complementary duplexes, respectively (SantaLucia and Hicks, 2004). As with any theoretical approach, the results of these equations should be used with caution. Some experiments may involve reagents (e.g. tetramethylammonium chloride) or conditions for which these equations are unsuitable. In these cases, only an empirical approach may achieve a satisfactory result.

1.2. Target sequences for detection of microbes

In the recent years, real-time quantitative PCR (qPCR) is considered as a method of choice for detection and quantification of microorganisms (Postollec et al., 2011). Also DNA sequencing (specific genes) has recently been applied in a clinical laboratory under routine conditions. Despite the sequencing technology now being automated, the accuracy and the interpretation of results remain difficult (Postollec et al., 2011; Woo et al., 2003; Christensen et al., 2005; Woo et al., 2008; Mignard and Flandrois, 2006). Nevertheless, currently, prokaryotes are mainly identified through DNA sequences and adequate selection of target sequence is a matter of question.

One of the most frequently used target sequences to detect prokaryotes is the bacterial ribosomal operon (16S rRNA, 23S rRNA and intergenic spacer (IGS) region). It is ubiquitous, contains both variable and highly conserved sequences, and usually results in sensitive detection due to its multicopy nature (Chakravorty et al., 2007). rRNA gene sequences for many species are also easily downloadable from public databases, which are rapidly growing and constantly being updated. While the discriminating power of 16S rRNA gene can be sufficient for some genera and species, it is not always enough to distinguish closely related species (Chakravorty et al., 2007). Therefore, other housekeeping genes have been studied, as well as functional genes involved in virulence or metabolism. For example, housekeeping genes encoding RNA polymerase, DNA helicase and DNA gyrase, have been used effectively as target sequences when detecting bacteria (Dahllof et al., 2000; Hu et al., 2011a; Thorsen et al., 2011). Several PCR-tests are based on pathogenic genes, e.g. shiga toxin-specific identification of *Escherichia coli* (Chui et al., 2010). However, methods detecting bacteria by protein coding genes frequently fail because of polymorphic target sequences or closely related species (Ziemer and Steadham, 2003; Balboa et al., 2011; Simões et al., 2011).

As plasmids are present with multiple copies in bacterial cell and contain sequences lacking extensive similarity to other sequences of host (antibiotics and heavy metal resistance, bacteriocins, restriction-modification systems), plasmidic sequences have been used for detecting bacteria (Davidson et al., 1996; Dougherty et al., 1998; Zhao et al., 2010; Irenge et al., 2010; Cevallos et al., 2008). However, as plasmids can be horizontally inherited between bacteria and may be dispensable for host survival, it may not be reliable to use plasmids for the identification of bacteria (Soda et al., 2008; Smillie et al., 2010).

1.3. Prokaryotic repetitive sequences

Repetitive sequences form only a small fraction of prokaryotic genomes. However, some distinct regions in bacterial and archaeal genomescan be considered as repetitive sequences, since they are present at least in two copies. Next are

introduced major classes of repetitive sequences also analyzed in the Results section of this thesis; starting with mobilome, followed with protein coding genes and rRNA genes and ending with non-coding repeats.

1.3.1. Mobilome

Mobile genetic elements (MGEs) are segments of DNA that encode enzymes and other proteins that mediate transfer of DNA within bacterial genomes (intracellular mobility) or between bacterial cells (intercellular mobility). The mobilome consists of bacteriophages, plasmids, transposable elements and genes that are often associated with them that regularly become passengers, such as restriction–modification (RM) and toxin–antitoxin (TA) systems (Koonin and Wolf, 2008).

Movement of DNA between bacterial cells can be conducted in three ways: transformation, conjugation and transduction. Transformation involves the transfer of cellular/naked DNA between closely related bacteria and is mediated by chromosomally encoded proteins found in some naturally transformable bacteria. In contrast, conjugation requires independently replicating genetic elements called conjugative plasmids or chromosomally integrated conjugative elements (ICEs), which include conjugative transposons. These genetic elements encode proteins that facilitate their own transfer and occasionally the transfer of other cellular DNA from the 'donor' plasmid-carrying cell to a recipient cell that lacks the plasmid or ICE. The process requires direct contact between the recipient and donor cell. Transduction is also a form of DNA transfer mediated by independently replicating bacterial viruses, called bacteriophages (or phages). At low frequency, bacteriophages can accidentally package segments of host DNA in their capsid and inject this DNA into a new host, where it can recombine with the cellular chromosome and be inherited (Frost et al., 2005).

Transposable elements

Transposable elements are mobile genetic elements that can move (transpose) from one site in the genome to a second site, or from one DNA molecule (that is, an infecting phage genome or a plasmid) to a second DNA molecule (the bacterial chromosome). They are the most abundant type of mobile genetic elements in the genomes of intracellular bacteria (Bordenstein and Reznikoff, 2005). Transposable elements are generally classified according to their transposition mechanism. Retrotransposons transpose via RNA intermediate (class I transposons), whereas DNA transposons transpose via DNA intermediate (class II transposons).

Three different types of prokaryotic retrotransposons (class I transposons) have been described: group II introns, retrons and diversity generating retroelements (DGR). *Group II introns*, the best characterized bacterial retrotransposons, are self-splicing introns that multiply via reverse transcription and

are capable of carrying out both self-splicing and retromobility reactions. They are the only type known to exhibit autonomous mobility (Simon et al., 2008). *Retrons* are genetic elements that produce multicopy single-stranded DNA covalently linked to RNA (msDNA which contains ssDNA part and ssRNA part) by a reverse transcriptase. The size of a retron varies between 1.3kbp to 3.0kbp. The functions of msDNA remain unknown (Lampson et al., 2005). *Diversity-generating retroelements* are a newly discovered family of genetic elements that confer selective advantages under certain conditions by introducing vast amounts of sequence diversity into target genes. Bacterial DGRs have in common a distinctive reverse transcriptase and two nearly identical repeat regions, the template region (TR) and variable region (VR). Central to this process is a reverse transcriptase-mediated exchange between the repeats, one serving as an donor template and the other as a recipient of variable sequence information (Doulatov et al., 2004). Approximately 25% of eubacterial genomes contain at least one retrotransposon constituting ≤1% from the genome (Simon et al., 2008).

Class II transposons are the most frequent and repetitive mobile genetic elements in bacterial genomes. Their size ranges from 0.7 to 3.5kbp and they generally encode only functions involved in their mobility. *Transposons* may carry unrelated genes like antibiotic resistance genes, catabolic genes and virulence determinants (Mahillon and Chandler, 1998). Transposable elements move from one genomic location to another by a process that is independent of sequence homology. The process follows one of the two pathways: conservative, to form simple insertions (also called cut-and-paste), and replicative, to form cointegrates (also known as copy-and-paste) (Treangen et al., 2009a).

Insertion sequences (IS) or Integrative and Conjugative elements (ICE) from the class II transposons are the most common autonomous mobile elements in bacterial genomes (Wozniak and Waldor, 2010). They are found in most eubacterial and archaeal genomes and they can comprise even up to 40% of prokaryotic genome. IS elements are up to 2kbp in size carrying one or two open reading frames (ORF) encoding transposase and the enzyme that catalyzes their movement, generally flanked by short terminal inverted repeats (Filee et al., 2007). Based on their organization and host range, insertion sequences are grouped into 20 families. Their transposition mechanism and specificity of target DNA sites varies enormously across different families of elements (Siguier et al., 2006b).

Miniature inverted transposable elements (MITE) are a subset of class II transposons. MITEs are short (generally <300bp) non-autonomous elements. They carry terminal inverted repeat sequences, and many of them also contain open reading frames (Delihas, 2008). Bacterial elements resemble archaeon and eukaryotic miniature inverted repeat transposable elements (Redder et al., 2001). MITEs are thought to derive from ISs by internal deletions and to be mobilized in trans by the transposase of their parental IS. The impact of MITE activity in the prokaryotic genome is potentially very high and their small size allows them to contribute in many ways to phenotypic variation, such as

generating new gene alleles and functions, or new regulatory signals for pre-existing genes (Delihas, 2008; Siguier et al., 2006a).

Plasmids

A plasmid is a collection of functional genetic modules organized into a stable, self-replicating entity, which is smaller than the cellular chromosome and usually does not contain genes required for essential cellular functions. Plasmid includes the essential genes that encode replicative functions and different accessory genes that encode processes distinct from those encoded by the bacterial chromosome. Several methods of horizontal movement of plasmids between bacterial cells occur: conjugation, where DNA is transported from the donor to the recipient cell; mobilizable methods, in which one plasmid parasitizes the self-directed transmission of another plasmid; transduction, in which a plasmid gets packaged in a phage particle; and transformation, in which cell lysis releases plasmid elements from the host bacterium. Plasmids also move vertically by transmission through dividing host cells (Davison, 1999; Frost et al., 2005).

Bacteriophages

Most of the bacterial genomes contain phage sequences, some of them can occupy up to 20% of the host genome. Phages are the most abundant and the most rapidly replicating life forms on earth, with enormous genetic diversity (Canchaya, et al., 2003). The genomes of phages can be composed of either single- or double-stranded DNA or RNA, and can range in size from a few to several 100 kb. Their characteristic essential genes comprise specific replicase genes, genes encoding phage components that 'hijack' the host cell replicative machinery, and genes encoding the proteins that package DNA into a protein coat (the capsid). Virulent bacteriophages lyse the host bacteria. Temperate bacteriophages have an alternative, quiescent, non-lytic growth mode called lysogeny. In most known cases of lysogeny, the phage genome integrates into the bacterial chromosome and replicates with it as a prophage; but in a few cases, the phage genome replicates autonomously (Lwoff, 1953) as a circular or linear plasmid (Canchaya et al., 2003). Environmental stimuli, such as DNA damaging agents, provoke a switch from quiescent to virulent replication that leads to cell lysis during which host cell DNA can be accidentally packaged and later injected into a new host in a process called transduction (Zinder and Lederberg, 1952). The ability to transduce host DNA seems to be limited to relatively large (50–100 kb) double-stranded DNA phages. The transduced chromosomal DNA must be able to recombine with the genome of the recipient host to survive (Frost et al., 2005).

1.3.2. Repeats associated with protein coding genes and rRNA genes

Protein domains can be divided into families or super-families by the fact that their members descended from a common ancestor. Protein domains are recurrent fragments with distinct structure, function, and evolutionary history. The ability to detect the evolutionary relationships of domains by sequence similarity is limited because they frequently diverge beyond the point where true relationships become unrecognizable (Chothia and Gough, 2009). Protein domains may occur alone, but are more frequently found in combination with other domains in multidomain proteins. The creation of new multidomain architectures through the shuffling of protein domains has been extensively studied. Domain repeats contain two or more domains from the same family in tandem. These repeats have a variety of binding properties and are involved in protein–protein interactions, as well as binding to other ligands such as DNA and RNA. (Bjorklund et al., 2006). Across a large variety of proteome species, the multi-domain protein universe is thin, but wide in respect to the distribution of domains in the proteome (Dokholyan et al., 2002). Although distinct protein domain families exist, different members of the same domain family may not have recognizable similarity according to their DNA bases. Thus they may be repetitive only in the sense of the amino acids.

Most genes are not unique but are part of larger families of related genes. Gene families originate by duplication of ancestral genes, after which they in turn have been duplicated. Increased duplication is responsible for increased genomic and phenotypic complexity. There is evidence that gene duplication is a contiguous and frequently occurring process. Large amounts of genomic data seem to suggest that many duplicates have been formed during some large-scale gene duplication events (Raes and Van de Peer, 2003), and therefore may have played a fundamental role in the evolution. In all three domains of life, large proportions of genes have been generated by gene duplication. It is suggested that these proportions are underestimates, because many duplicated genes have diverged so that sequence similarity has been lost (Zhang, 2003). Gene duplication can result from unequal crossing-over, retroposition or chromosomal duplication. The duplicated region may contain part of a gene, an entire gene or several genes. A gene duplicate will ultimately suffer one of three fates: one copy will be silenced by degenerative mutation (becomes a pseudogene), one copy will evolve a new beneficial function (neofunctionalization) that is permanently preserved in the population, or both copies may be reciprocally preserved through the fixation of complementary loss-of-subfunction mutations (subfunctionalization), which result in partitioning of the tasks of the ancestral gene (Lynch et al., 2001). Pseudogenes will be either deleted from the genome or become so divergent from the parental genes that they are no longer identifiable. Furthermore, the presence of duplicate genes is sometimes beneficial, simply because extra amounts of protein or RNA products are provided. This

applies mainly to strongly expressed genes whose products are in high demand, such as rRNAs or housekeeping genes (Zhang, 2003; Gevers et al., 2004).

Gevers et al analyzed prokaryotic paralogs restricted to one strain (genes without orthologs) in 106 prokaryotic genomes. They reported that nearly all genomes contained strain-specific sequences but in different numbers. Nearly half of the proteins of strain-specific genes (~40%) are annotated as 'protein of unknown function' (Gevers et al., 2004). It is suggested that these genes have been derived either *de novo* in the current strain or that they diverged to an extent where no sequence similarity to sequences in other strains could be detected (Jordan et al., 2001).

1.3.3. Non-coding repeats

Specific families of repeated DNA elements have been found in prokaryotic genomes in intergenic regions, but their origin and the function are not well understood. They are thought to play a role in the evolution of the genome, and in genome architecture, structure and plasticity (Treangen et al., 2009a). As follows, several more extensively studied types of short intergenic repetitive sequences are introduced.

The *repetitive extragenic palindromic* (REP) sequences, also called palindromic units (PU), are one of the best-described classes of intergenic repeats. Their length ranges between 20–60bp, they possess an imperfect palindromic core, and they occur hundreds of times within a genome (Stern et al., 1984; Higgins et al., 1988). REP sequences may form stem-loop structure and are probably transcribed. Functions of REP have been associated with the regulation of gene expression. While often existing as singlets, REPs also form a range of complex higher order structures termed BIMEs (bacterial interspersed mosaic elements). In these mosaic elements, REPs are combined with different short conserved sequence motifs forming different classes of BIME. BIMEs vary in length between 40–500bp (Gilson et al., 1991).

Enterobacterial repetitive intergenic consensus elements (ERIC), also called intergenic repeat units (IRU), are of 69–127bp in length and are probably transcribed (Hulton et al., 1991; Sharples and Lloyd, 1990). The function of ERIC sequences is unclear. They may be involved in the regulation of mRNA stability (Delihas, 2008).

Clustered regularly interspaced short palindromic repeats (CRISPR) are hypervariable genetic loci widely distributed in bacterial (~40%) and in most archaeal (~90%) genomes (Jansen et al., 2002; Sorek et al., 2008; van der Oost et al., 2009). They are composed of direct repeats, repeated up to 250 times, ranging in size from 24 to 47 nt that are separated by similarly sized (21–72bp) non-repetitive spacers. In the large majority of cases, the direct repeats are highly conserved while the spacers are very diverse within a given locus, even among strains of the same species. Recent studies have established that CRISPR provide acquired resistance against foreign DNA, such as plasmids and phages

(Barrangou et al., 2007). CRISPR sequences are located adjacent to cas-genes (CRISPR associated genes), which may be involved in the propagation and function of these repeats. *Cas*-genes encode large and heterogeneous protein families, like nucleases, helicases, polymerases and polynucleotide-binding nucleotides (Haft et al., 2005). CRISPR with cas-genes forms the CRISPR/Cas-system. For example, in CRISPR/Cas systems *cas1* and *cas2* six core cas-genes and universal markers have been identified. Most of the repeat sequences are partially palindromic, with the possibility of forming a stable conserved secondary structure. The exact mechanism behind the CRISPR/Cas system providing resistance against foreign DNA is still unclear. Some Cas proteins are involved in the acquisition of novel spacers from foreign DNA, whilst others provide CRISPR-encoded phage resistance and interfere with invasive genetic elements. The principles behind invasive element recognition, novel repeat manufacturing, and spacer selection and integration into the CRISPR locus remain uncharacterized (Horvath and Barrangou, 2010).

1.4. Current *ab initio* methods for finding long repetitive sequences

Two sorts of programs exist for finding long repetitive sequences, programs using predefined database of repetitive sequences and programs that do not need prior knowledge. Discovery of new repetitive sequences from genomes remains a challenging computational problem due to the different characteristics of repeats, along with the necessity of allowing substitutions and gaps in the alignments (Treangen et al., 2009a). Currently, many programs are available for finding new repetitive sequences from complete genomes. Here we discuss some programs that search for them solely on the basis of their repetitive nature (not the basis of structure) from assembled genomes. These programs differ from each other mainly by their performance – specificity, sensitivity, running time, need for memory, etc. (Saha et al., 2008). The widely used Reputer (Kurtz et al., 2001) is non-heuristic on a suffix-tree based computer program. A repeat is viewed like a substring of length k that occurs more than once in a sequence. After finding exact repeats (in linear space and time), it uses them as seeds to construct degenerate repeats allowing for mismatches, insertions and deletions. It guarantees to find all degenerate repeats. The heuristic program RECON (Bao and Eddy, 2002) is based on an approach of single linkage clustering of local pairwise alignments between genomic sequences. To obtain pairwise local alignments between sequences, RECON uses WU-BLASTN (Gish, W unpublished). Thereafter a graph is generated in which the vertices correspond to repeat elements and edges connect elements with a specified ratio of overlap. Elements with a high degree of overlap are assumed to belong to the same repeat family, whereas those with less similarity are assumed to correspond to related families. The boundaries of a repeat family are identified through aggregation of end-points. Repeatoire (Treangen et al., 2009b), EulerAlign

(Zhang and Waterman, 2005) and RepeatScout (Price et al., 2005) construct repeat families directly via local multiple alignment. These programs take heuristic approaches. Repeatoire is a program for finding interspersed repeats from bacterial genomes. It finds multi-matching seeds, aligns them and then performs gapped extension. The EulerAlign is based on graphs theory, where DNA sequence is broken into overlapping segments to construct a Bruijn graph. Cycles are then resolved from the graph to find consensus alignments. The computation time of EulerAlign is approximately linear with respect to the total size of data. RepeatScout builds a library of high-frequency k-mers and retrieves substrings of the input sequence containing a specific k-mer. A penalty-based local alignment of the substrings is used to extend the k-mer and generate a consensus sequence for each repeat family (Saha et al., 2008).

Currently, different programs are combined to discover repetitive sequences from newly sequenced genome (in case of eukaryotic species). Known repeats are identified via comparisons of query sequences with those in curated repeat libraries. For example, RepeatMasker (Smit, AFA, Hubley, R & Green, P. *RepeatMasker* 1996–2010, http://www.repeatmasker.org) is the most widely used tool in repeat discovery and classification. It is frequently used as a first step when an uncharacterized genome is searched for repeats. It uses a well-maintained pre-defined repeat library containing repeats for diverse eukaryotic organisms. However, as repeat families are largely species-specific, when a new genome is analyzed, a new repeat library should be generated. Various *ab initio* computer programs are used for large genomes to find novel repeats. For example, annotating repetitive sequences from genome of the mushroom, *Schizophyllum commune,* the repeat finding programs RepeatScout, Repeat-Masker and RECON have been used (Ohm et al., 2010). Also, RepeatMasker, BLAST, RepeatScout and RECON are used in many other newly sequenced genomes for repeat discovery (International Aphid Genomics Consortium, 2010; Schmutz et al., 2010; Chan et al., 2010; Hu et al., 2011b).

Though many previously described methods can be also applied for finding (species-specific) repetitive sequences from bacterial and archaeal species they all have some drawbacks. For example, Reputer outputs only pairwise relation-ship between instances of repeat, which requires exhaustive parsing of Reputer results by user. Also, applying Reputer one can define similarity parameter which is the maximum number of positions at which two copies of a particular repeat may differ. This forces user to find repeats with fixed number of differences in case of variable repeat lengths. Futher, the main disadvantages of RECON are its speed and inconvenient execution of program (first user has to run BLAST, thereafter user has to convert output to other format and finally user can run RECON). Although RepeatScout and EulerAlign are flexible and fast, in some cases they may not be preferred for finding repeats. For example, in case of finding species-specific repetitive sequences from bacterial and archaeal genomes, implementation of new method over exhaustive scripting that is required to find species-specific regions from found repeats, may be preferred.

2. RESULTS AND DISCUSSION

2.1. Aims of the study

The main goal of this work is to improve the sensitivity of PCR, especially in the case of detection of prokaryotic species from medical or environmental samples. PCR is still a widely used method for detecting prokaryotic organisms in many fields. However, its sensitivity is not always satisfactory, which is crucial in, e.g. clinical diagnostics and food pathology.

More constructively the aims of current work are:

a. to improve widely used PCR primer design program Primer3 (Rozen and Skaletsky, 2000) with more precise melting temperature formula. The average difference between the experimental and predicted melting temperature by old Primer3 (version 1.0 and older) in case of oligos 15–30 nt is 11.7^0C.
b. to develop a method for finding clade-specific repetitive sequences that can be used to detect prokaryotic species with PCR, to analyze the occurrence of species-specific repetitive sequences in prokaryotic species, and to test experimentally species-specific repetitive sequences in clinical diagnostics. Most widely applied marker gene has been 16S rDNA because it has universal and strong phylogenetic signal. However, 16S rDNA sequences cannot always distinguish closely related species or are too variable for distinguishing more divergent species. Species-specific repetitive sequences provide an alternative target sequence for more sensitive PCR detection.
c. to characterize long prokaryotic species-specific repetitive sequences and to find possible roles in the genomes for these repeats. Prokaryotic genomes are compact without excessive repetitive sequences. Most of the repetitive sequences are thought to be mobile genetic elements, which are common in prokaryotic species. However, no large scale analyses of species-specific repetitiveness of prokaryotic genomes have been conducted.

2.2. Improvements of primer design program Primer3 (Ref I)

One field in which our workgroup has been involved for several years is PCR primer design. Although currently the sequencing of DNA is rapidly evolving and increasingly being applied, the relative importance of PCR in practical fields remains the same. The need to update primer design program Primer3 arose from the project where a design of large number of PCR primers was required (Ref II).

As primer design is not a trivial process, it should be performed with the computer program, several of which exist, but one of the most widely used freeware program is Primer3. Until 2006, Primer3 code (version 1.0 and older, further referred as "*old*") was maintained mainly by one scientist, its creator - Steve Rozen. After suggesting to Steve Rozen that improvements of melting temperature calculations should be added to the main code of Primer3 (version 1.1 and newer, further referred as "*new*"), it became a major open source project. Thus scientists from both fields, biological and computational, began to improve the main code of Primer3.

Although the formula, for calculating the melting temperature of PCR primers and oligos in old Primer3, is based on the nearest-neighbour (NN) thermodynamics, the exact principle of the formula and the values of thermodynamic parameters used by this formula do not give the correct value of melting temperature. At the time when these formula and thermodynamic values were implemented into the old Primer3 code, they were one of the most precise ones (Breslauer et al., 1986; Rychlik et al., 1990). After SantaLucia JR. published a paper about DNA nearest-neighbours in 1998, it could be ascertained that the results of the latter were more precise (Koressaar et al., 2007). The approach for calculating the melting temperature of short duplexes implemented in old Primer3 has three main shortcomings. The approach implemented in old Primer3 is suggested by Rychlik et al (Rychlik et al., 1990) and it is based on three papers publishing the NN formula (Borer et al., 1974), the table of thermodynamic parameters (Breslauer et al., 1986), and the salt correction formula (Schildkraut and Lifson, 1965). First, the NN formula of Borer et al. assumes that helix initiation enthalpy and entropy are equal to zero. In the formation of a duplex, unfavourable entropy associated with the loss of translational freedom upon formation of the first hydrogen bonded base-pair (i.e. the initiation free energy) must be considered (SantaLucia et al., 1996). Second, the outdated table of thermodynamic parameters was used to calculate melting temperature. The table of thermodynamic parameters was published in 1986 by Breslauer et al (Breslauer et al., 1986). Several new and improved sets of NN parameters have been published (SantaLucia, 1998). The third shortcoming is associated with the salt correction component. As thermodynamic parameters are measured in 1M Na^+, it is necessary to use correction in case PCR buffers with different salt concentration when calculating the melting temperature of a short duplex. Different monovalent and different divalent

cations can have a similar effect on duplex stability. However, divalent cations stabilize DNA duplexes significantly more than the same concentration of monovalent cations (Owczarzy et al., 2008). The nearest-neighbour approach of Rychlik et al. uses a very trivial salt correction component that was suggested by Schildkraut and Lifson (Schildkraut and Lifson, 1965). The salt correction formula was developed to model the melting behaviour of a specific bacterial genomic DNA sequence within a narrow range of sodium ion concentrations. Nevertheless, it is routinely applied to model the behavior of short DNA oligomers of any base composition in a wider range of ionic environments, but with no evidence to support that such generalizations are valid (Owczarzy et al., 2008). Another problem with the salt correction component in the approach by Rychlik et al. is that it enables the use of only monovalent cations in the melting temperature calculation. The formula with salt correction component used by Primer3 before improvements by Rychlik et al. is:

$$\mathrm{Tm} = \frac{\Delta H}{\Delta S + R\ln(c/4)} - 273.15 + 16.6\log[K^+]$$

where ΔH and ΔS are the enthalpy and entropy for helix formation, respectively, R is the molar gas constant (1.987 cal/C° x mol), and c is the total molar concentration of hybridizing molecules where oligonucleotides are not self-complementary (Rychlik et al., 1990).

Improvements introduced to Primer3 comprise:

a. An update of the table of thermodynamic parameters (Table 2 from SantaLucia, Proc. Natl. Acad. Sci, 1998), which is required to calculate primers/oligos melting temperature based on the nearest-neighbour model (SantaLucia, 1998).
b. The update of nearest-neighbour model for calculating melting temperature of short duplexes (Ref I formula I) (SantaLucia, 1998). This model also contains the duplex formation initiation parameters and symmetry penalty for self-complementary duplexes.
c. Implementation of two salt correction formulas with the new possibility of using divalent cations in the PCR buffer (Ref I formulae II and III) (Owczarzy et al., 2004; SantaLucia, 1998)

As is shown in Ref I, for primers of typical length (15–30 nucleotides) the average differences between the experimental and predicted Tm is 1.37 and 11.7 °C for new and old Primer3 respectively. However, primers with size between 15–30 nucleotides have melting temperature between 35.5°C and 87.7°C (in our dataset) which is not a common PCR melting temperature range. Thus we conducted an additional analysis to clarify the accuracy of new melting temperature formula in Primer3 using oligos which experimental melting temperature falls to common PCR melting temperature range (45°C to 65°C). In this case, the average difference between the experimental and predicted melting temperature by old Primer3 is 9.8°C (lower and upper quartile 6.5°C

and 12.1°C, respectively). The average difference between the experimental and predicted melting temperature by new Primer3 (calculations are conducted using SantaLucia's salt correction formula) is 2.1°C (lower and upper quartile 0.7°C and 2.5°C, respectively). The experimental melting temperatures and corresponding PCR buffer conditions were retrieved from literature (Owczarzy et al., 2004; Owczarzy et al., 2008) and include 487 different measurements and 89 different oligonucleotides with length ≤30nt.

2.3. Prokaryotic species-specific repetitive sequences for species-specific primer design (Ref II)

Nowadays, PCR-based methods, in particular quantitative PCR (multiplex qPCR), are used predominantly to detect, identify and quantify either pathogens or valuable species (i.e. fermenting microbes or probiotics in food microbiology). Since the detection of new (pathogenic or beneficial) bacteria is constantly needed, the primer design should be quick and produce primers that detect all preferred genomes with high confidence and do not detect undesired ones. Also, new primers replacing existing ones are frequently required (i.e. new isolates of a species have appeared that cannot be detected with existing primers).

The concentration of target genomes in biological sample is frequently low causing false negative PCR results. Therefore approaches allowing the design of more sensitive primers are needed. It can be presumed (based on repetitive 16S rRNA genes) that if PCR primers have more than one annealing site in the molecule, then the probability that primer hybridizes to its binding site is higher and therefore the sensitivity of PCR is higher.

2.3.1. Methodology for species-specific primer design

Method for finding species-specific repeats

We have developed a method for finding species-specific repeats with the aim of using these sequences in PCR primer design. This method can also be applied to find repeats specific to some other taxonomic unit, e.g. some specific pathogenic group of strains or genera.

Further, the method of finding repeats is described briefly, based on the example of finding species-specific repeat. First, the complete genomic sequences for species of interest are needed (*target genomes*) to find repetitive sequences. If more than one genome is sequenced from a species, then a species-specific repeat must occur in all these genomes. Basically, for finding species-specific repeats, the genomic sequence of a strain is split into sequences with predefined length and with predefined overlap length. Thereafter, the repetitiveness of a split and the species-specificity are both checked by the homology search tool BLAST. The decision if a sequence is repetitive is based

on the match length and the BLAST bit score. The specificity of the sequence is decided by the length of a match and the BLAST bit score of a match found from a non-target prokaryotic genome; however, somewhat looser cut-off values are used to eliminate all possible unspecific sequences. Finally, overlapping repeats are joined into one repeat copy. A schematic representation is given in Figure 1 (Ref II).

Method for species-specific primer design

To use species-specific repetitive sequences as PCR target sequences, we first align all copies of a repeat with each other by multiple sequence alignment tool CLUSTALW. Thereafter, we construct consensus sequence from multiple sequence alignment where non-canonical nucleotides, variable nucleotides and *indels* are marked with the symbol 'N'. Next, the consensus sequence is used as the PCR target sequence to design a candidate primer with Primer3. No 'N' symbols are allowed in the primer sequence. Finally, the specificity of candidate primers is checked. Alternative PCR products (with predefined length) are predicted for every candidate primer pair from all non-target genomes. The candidate primers binding sites are searched with a program which is based on thermodynamic alignment (FastaGrep). As a cut-off measure (whether primer binds to alternative site or not), free energy value of duplex (with mismatches) between candidate primer and alternative site, is used. If the hybridization is equal to or is more stable than the free energy of 12 nucleotides of full duplex of primers 3'end, then the binding site is considered in the calculation of alternative PCR products. All primer pairs, to which alternative PCR products are predicted, are rejected. The illustrative scheme is presented in the Figure 3 (RefII).

We have also implemented a web interface for the method to enable the design of PCR primers for detecting a specific group of genomes. The web service is called MultiMPrimer3 and can be found at http://bioinfo.ut.ee/multimprimer3/.

2.3.2. Occurrence of species-specific repetitive sequences in 30 randomly chosen genomes

We randomly chose 30 genomes (from 508 completely sequenced genomes), to get an overview of prokaryotic species-specific repetitiveness. We analyzed how many of these contain species-specific repetitive sequences (≥80% of similarity between copies of particular repeat was used as similarity cutoff), what is the size range of repeats per different species and how many copies per different repeats can be found. All species-specific repetitive sequences with the length not less than 100, 300 and 1000 bp were searched. The results are shown in the Table 1 (Ref II). All species analyzed contain at least one species-specific repeat with a length of at least 300 nt. We also briefly checked whether strain-specific repeats were present. The strain-specific repeats could be found over half of the 17 strains (three species) analyzed. Thus, at least in some cases,

the smallest taxonomic unit that could be identified with the unit-specific repetitive DNA sequence is strain.

2.3.3. Experimental testing of PCR primers designed on species-specific repeats

To test experimentally the hypothesis that species-specific repeats could increase the sensitivity of PCR, we chose four bacterial pathogenic species and one fungal genome (*Helicobacter pylori, Listeria monocytogenes, Mycoplasma genitalium, Neisseria gonorrhoeae, Candida albicans*, respectively). More precisely, we selected 132 different primer pairs designed, whether on species-specific repeats with 2–16 copies or on species-specific non-repeated regions. The gel electrophoresis band intensities were visually assessed on the scale 1 to 5, where 5 represents the highest intensity. We found that the correlation between the number of copies per repeat and the gel electrophoresis band intensity was statistically significant ($P<0.0001$ of F-statistic).

2.4. Characterization of species-specific repetitive sequences (Ref III)

Motivated by the previous paper, we decided to characterize the repetitive sequences of prokaryotic genomes. First, we were concerned about the sequences behind the species-specific repeats. Common knowledge is that repetitive elements in prokaryotic genomes are mobile genetic elements, 16S rRNA genes and a few conserved duplicated genes. No report about large-scale study of species-specific prokaryotic repeats seems to have been published.

We considered it important to analyze prokaryotic species-specific repetitive sequences for two reasons. The first is a general characterization of prokaryotic species through species-specific repeats. The second is the need to comprehend the sequences under the species-specific repeats if one intends to use them as PCR target sequences. The main questions are how common are species-specific repeats and what categories/types of species-specific sequences can be distinguished.

We searched repetitive sequences from 613 different prokaryotic species (876 completely sequenced genomes) with the method described in Ref II. Only a brief description of the parameter values used will be provided. The genomic sequence of a target genome was split into 100bp long sequences. The adjacent splits were allowed to overlap with each other 50bp. The repetitiveness and species-specificity of a split was checked with the homology search algorithm BLAST. A split was considered as candidate repeat if the length of BLAST match was between 85–115bp and the identity between the matching region and the split was >80% (this percentage of similarity preserves sufficiently conserved regions in consensus sequence, generated from multiple sequence align-

ment of copies of particular repeat, to design PCR primers). A split was considered as non-species-specific if the length of a BLAST match in any of the non-target genomes had length at least 50bp and the BLAST identity was >60%.

Main findings of this paper are as follows.

1. Almost all species analyzed contain repetitive sequences. We could not identify repeats in just three species which had extremely small genomes.
2. The fraction of genome covered by repeats is small (the median value of repeat coverage 1.8% and the median value of species-specific repeat coverage is 1% of the genome). Only 12 bacterial species had the coverage of repetitive sequences from the genome >7%.
3. Most of the species contain species-specific repeats. We were unable to detect species-specific repeats in 20 species.
4. A large fraction of the species-specific repeats were associated with protein-coding genes. After classification of species-specific repeats we concluded that 64% of repeats were associated with protein coding genes, 14% of repeats with mobile genetic elements, 13% with non-coding short repeats, 3% with RNA genes and 6% of repeats with unknown origin.
5. The functional analysis revealed that different repeat classes appear in different species. In some species, phage-related sequences are in prevalence, in others only rDNA related repeats appear, or, in the third, one large duplication containing different protein coding genes are found.

CONCLUSIONS

All three papers included to thesis contribute to successful identification of microbial organisms. As micro-organisms are mostly identified with the aid of DNA sequences, it is important to develop DNA based identification methods, which among other things comprise PCR primer design and finding alternative target sequences. In this thesis, several improvements of prokaryotic PCR primer design have been suggested. This thesis also provides an exhaustive study of prokaryotic species-specific repetitive sequences which comprises the characterization of such sequences and raising assurance when applying them as PCR targets.

The relevance of this thesis can be concluded as follows:

1. Determination of annealing temperature is a critical step in PCR. As annealing temperature is derived from the calculated melting temperature, its estimation must be accurate. The improvements, we introduced to widely used primer design program Primer3, enable the calculation of primer melting temperature more precisely. The average difference of the experimental and estimated melting temperature of short oligos (15–30 nucleotides, which experimental melting temperature is between 35°C and 88°C) of old Primer3 (version 1.0 and older) and the new Primer3 (version 1.1 and newer) are 11.7°C and 1.37°C, respectively.
2. We proved statistical significance of the hypothesis that the higher number of copies per species-specific repeats result to risen sensitivity of PCR. Also the methodology for finding prokaryotic species-specific repeats with species-specific PCR primer design has been provided.
3. Large scale characterization of prokaryotic species-specific repeats denotes that most of prokaryotic genomes contain species-specific repeats. Functional analysis indicates that most of the species-specific repeats are associated with protein coding sequences and in different species variable functions could be linked to species-specific repeats.

REFERENCES

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

Balboa, S., Doce, A., Dieguez, A.L. and Romalde, J.L. (2011) Evaluation of different species-specific PCR protocols for the detection of Vibrio tapetis. *J.Invertebr.Pathol.,* 108, 85–91.

Bao, Z. and Eddy, S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.,* 12, 1269–1276.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* 315, 1709–1712.

Bjorklund, A.K., Ekman, D. and Elofsson, A. (2006) Expansion of protein domain repeats. *PLoS Comput.Biol.,* 2, e114.

Bordenstein, S.R. and Reznikoff, W.S. (2005) Mobile DNA in obligate intracellular bacteria. *Nat.Rev.Microbiol.,* 3, 688–699.

Borer, P.N., Dengler, B., Tinoco, I.,Jr and Uhlenbeck, O.C. (1974) Stability of ribonucleic acid double-stranded helices. *J.Mol.Biol.,* 86, 843–853.

Breslauer, K.J., Frank, R., Blocker, H. and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc.Natl.Acad.Sci.U.S.A.,* 83, 3746–3750.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L. and Brussow, H. (2003) Phage as agents of lateral gene transfer. *Curr.Opin.Microbiol.,* 6, 417–424.

Cevallos, M.A., Cervantes-Rivera, R. and Gutierrez-Rios, R.M. (2008) The repABC plasmid family. *Plasmid,* 60, 19–37.

Chakravorty, S., Helb, D., Burday, M., Connell, N. and Alland, D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J.Microbiol.Methods,* 69, 330–339.

Chamberlain, J.S., Gibbs, R.A., Ranier, J.E., Nguyen, P.N. and Caskey, C.T. (1988) Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res.,* 16, 11141–11156.

Chan, A.P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K.M., Redman, J., Chen, G.and et al. (2010) Draft genome sequence of the oilseed species Ricinus communis. *Nat.Biotechnol.,* 28, 951–956.

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.

Cheng, Y., Korolev, N. and Nordenskiold, L. (2006) Similarities and differences in interaction of K+ and Na+ with condensed ordered DNA. A molecular dynamics computer simulation study. *Nucleic Acids Res.,* 34, 686–696.

Chothia, C. and Gough, J. (2009) Genomic and structural aspects of protein evolution. *Biochem.J.,* 419, 15–28.

Chou, Q., Russell, M., Birch, D.E., Raymond, J. and Bloch, W. (1992) Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. *Nucleic Acids Res.,* 20, 1717–1723.

Christensen, J.J., Andresen, K., Justesen, T. and Kemp, M. (2005) Ribosomal DNA sequencing: experiences from use in the Danish National Reference Laboratory for Identification of Bacteria. *APMIS,* 113, 621–628.

Chui, L., Couturier, M.R., Chiu, T., Wang, G., Olson, A.B., McDonald, R.R., Antonishyn, N.A., Horsman, G. and Gilmour, M.W. (2010) Comparison of Shiga

toxin-producing Escherichia coli detection methods using clinical stool samples. *J.Mol.Diagn.,* 12, 469–475.
Dahllof, I., Baillie, H. and Kjelleberg, S. (2000) rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl.Environ.Microbiol.,* 66, 3376–3380.
Davidson, B.E., Kordias, N., Dobos, M. and Hillier, A.J. (1996) Genomic organization of lactic acid bacteria. *Antonie Van Leeuwenhoek,* 70, 161–183.
Davison, J. (1999) Genetic exchange between bacteria in the environment. *Plasmid,* 42, 73–91.
Delihas, N. (2008) Small mobile sequences in bacteria display diverse structure/ function motifs. *Mol.Microbiol.,* 67, 475–481.
Dokholyan, N.V., Shakhnovich, B. and Shakhnovich, E.I. (2002) Expanding protein universe and its origin from the biological Big Bang. *Proc.Natl.Acad.Sci.U.S.A.,* 99, 14132–14136.
Dougherty, B.A., Hill, C., Weidman, J.F., Richardson, D.R., Venter, J.C. and Ross, R.P. (1998) Sequence and analysis of the 60 kb conjugative, bacteriocin-producing plasmid pMRC01 from Lactococcus lactis DPC3147. *Mol.Microbiol.,* 29, 1029–1038.
Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., Simons, R.W., Zimmerly, S. and Miller, J.F. (2004) Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature,* 431, 476–481.
Filee, J., Siguier, P. and Chandler, M. (2007) Insertion sequence diversity in archaea. *Microbiol.Mol.Biol.Rev.,* 71, 121–157.
Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc.Natl.Acad.Sci.U.S.A.,* 83, 9373–9377.
Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat.Rev.Microbiol.,* 3, 722–732.
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.
Gevers, D., Vandepoele, K., Simillon, C. and Van de Peer, Y. (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.,* 12, 148–154.
Gilson, E., Saurin, W., Perrin, D., Bachellier, S. and Hofnung, M. (1991) Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.,* 19, 1375–1383.
Haas, S., Vingron, M., Poustka, A. and Wiemann, S. (1998) Primer design for large scale sequencing. *Nucleic Acids Res.,* 26, 3006–3012.
Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput.Biol.,* 1, e60.
Heid, C.A., Stevens, J., Livak, K.J. and Williams, P.M. (1996) Real time quantitative PCR. *Genome Res.,* 6, 986–994.
Higgins, C.F., McLaren, R.S. and Newbury, S.F. (1988) Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. *Gene,* 72, 3–14.
Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science,* 327, 167–170.

Hu, Q., Tu, J., Han, X., Zhu, Y., Ding, C. and Yu, S. (2011a) Development of multiplex PCR assay for rapid detection of Riemerella anatipestifer, Escherichia coli, and Salmonella enterica simultaneously from ducks. *J.Microbiol.Methods,* 87, 64–69.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H.and et al. (2011b) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat.Genet.,* 43, 476–481.

Hulton, C.S., Higgins, C.F. and Sharp, P.M. (1991) ERIC sequences: a novel family of repetitive elements in the genomes of Escherichia coli, Salmonella typhimurium and other enterobacteria. *Mol.Microbiol.,* 5, 825–834.

International Aphid Genomics Consortium. (2010) Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol.,* 8, e1000313.

Irenge, L.M., Durant, J.F., Tomaso, H., Pilo, P., Olsen, J.S., Ramisse, V., Mahillon, J. and Gala, J.L. (2010) Development and validation of a real-time quantitative PCR assay for rapid identification of Bacillus anthracis in environmental samples. *Appl.Microbiol.Biotechnol.,* 88, 1179–1192.

Jansen, R., Embden, J.D., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol.Microbiol.,* 43, 1565–1575.

Johnson, J.R. (2000) Development of polymerase chain reaction-based assays for bacterial gene detection. *J.Microbiol.Methods,* 41, 201–209.

Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. and Koonin, E.V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.,* 11, 555–565.

Kampke, T., Kieninger, M. and Mecklenburg, M. (2001) Efficient primer design algorithms. *Bioinformatics,* 17, 214–225.

Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.,* 36, 6688–6719.

Koressaar, T. and Remm, M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics,* 23, 1289–1291.

Koressaar, T., Joers, K. and Remm, M. (2009) Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms. *Bioinformatics,* 25, 1349–1355.

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.

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.

Lampson, B.C., Inouye, M. and Inouye, S. (2005) Retrons, msDNA, and the bacterial genome. *Cytogenet.Genome Res.,* 110, 491–499.

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.

Lwoff, A. (1953) Lysogeny. *Bacteriol.Rev.,* 17, 269–337.

Lynch, M., O'Hely, M., Walsh, B. and Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics,* 159, 1789–1804.

Lyubartsev, A.P. and Laaksonen, A. (1998) Molecular dynamics simulations of DNA in solutions with different counter-ions. *J.Biomol.Struct.Dyn.,* 16, 579–592.
Mahillon, J. and Chandler, M. (1998) Insertion sequences. *Microbiol.Mol.Biol.Rev.,* 62, 725–774.
Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol.Biol.,* 453, 3–31.
Marky, L.A. and Breslauer, K.J. (1982) Calorimetric determination of base-stacking enthalpies in double-helical DNA molecules. *Biopolymers,* 21, 2185–2194.
Mignard, S. and Flandrois, J.P. (2006) 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *J.Microbiol.Methods,* 67, 574–581.
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.
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.
Nakano, S., Fujimoto, M., Hara, H. and Sugimoto, N. (1999) Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Res.,* 27, 2957–2965.
O'Connor, L. and Glynn, B. (2010) Recent advances in the development of nucleic acid diagnostics. *Expert Rev.Med.Devices,* 7, 529–539.
Ohm, R.A., de Jong, J.F., Lugones, L.G., Aerts, A., Kothe, E., Stajich, J.E., de Vries, R.P., Record, E., Levasseur, A., Baker, S.E.and et al. (2010) Genome sequence of the model mushroom Schizophyllum commune. *Nat.Biotechnol.,* 28, 957–963.
Onodera, K. and Melcher, U. (2004) Selection for 3' end triplets for polymerase chain reaction primers. *Mol.Cell.Probes,* 18, 369–372.
Owczarzy, R., Moreira, B.G., You, Y., Behlke, M.A. and Walder, J.A. (2008) Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry,* 47, 5336–5353.
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.
Owczarzy, R., You, Y., Moreira, B.G., Manthey, J.A., Huang, L., Behlke, M.A. and Walder, J.A. (2004) Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry,* 43, 3537–3554.
Panjkovich, A. and Melo, F. (2005) Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics,* 21, 711–722.
Postollec, F., Falentin, H., Pavan, S., Combrisson, J. and Sohier, D. (2011) Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol.,* 28, 848–861.
Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics,* 21 Suppl 1, i351–8.
Raes, J. and Van de Peer, Y. (2003) Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl.Bioinformatics,* 2, 91–101.
Redder, P., She, Q. and Garrett, R.A. (2001) Non-autonomous mobile elements in the crenarchaeon Sulfolobus solfataricus. *J.Mol.Biol.,* 306, 1–6.
Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol.Biol.,* 132, 365–386.

Rychlik, W., Spencer, W.J. and Rhoads, R.E. (1990) Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.,* 18, 6409–6412.
Saha, S., Bridges, S., Magbanua, Z.V. and Peterson, D.G. (2008) Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.,* 36, 2284–2294.
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.
SantaLucia, J.,Jr and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu.Rev.Biophys.Biomol.Struct.,* 33, 415–440.
SantaLucia, J.,Jr, Allawi, H.T. and Seneviratne, P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry,* 35, 3555–3562.
Schildkraut, C. and Lifson, S. (1965) Dependence of the melting temperature of DNA on salt concentration. *Biopolymers,* 3, 195–208.
Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J.and et al. (2010) Genome sequence of the palaeo-polyploid soybean. *Nature,* 463, 178–183.
Sharples, G.J. and Lloyd, R.G. (1990) A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes. *Nucleic Acids Res.,* 18, 6503–6508.
Siguier, P., Filee, J. and Chandler, M. (2006a) Insertion sequences in prokaryotic genomes. *Curr.Opin.Microbiol.,* 9, 526–531.
Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. and Chandler, M. (2006b) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.,* 34, D32–6.
Simon, D.M., Clarke, N.A., McNeil, B.A., Johnson, I., Pantuso, D., Dai, L., Chai, D. and Zimmerly, S. (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA,* 14, 1704–1713.
Simões, P.M., Mialdea, G., Reiss, D., Sagot, M.-. and Charlat, S. (2011) Wolbachia detection: an assessment of standard PCR Protocols. *Molecular Ecology Resources,* 11, 567–572.
Singh, V.K., Govindarajan, R., Naik, S., and Kumar, A. (2000) The Effect of Hairpin Structure on PCR Amplification Efficiency. *Molecular Biology Today,* 1, 67–69.
Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P. and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol.Mol.Biol.Rev.,* 74, 434–452.
Smith, C.J. and Osborn, A.M. (2009) Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol.Ecol.,* 67, 6–20.
Soda, S., Otsuki, H., Inoue, D., Tsutsui, H., Sei, K. and Ike, M. (2008) Transfer of antibiotic multiresistant plasmid RP4 from escherichia coli to activated sludge bacteria. *Journal of Bioscience and Bioengineering,* 106, 292–296.
Sorek, R., Kunin, V. and Hugenholtz, P. (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat.Rev. Microbiol.,* 6, 181–186.
Srivastava, G.P., Guo, J., Shi, H. and Xu, D. (2008) PRIMEGENS-v2: genome-wide primer design for analyzing DNA methylation patterns of CpG islands. *Bioinformatics,* 24, 1837–1842.
Stern, M.J., Ames, G.F., Smith, N.H., Robinson, E.C. and Higgins, C.F. (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell,* 37, 1015–1026.
Sugimoto, N., Nakano, S., Yoneyama, M. and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.,* 24, 4501–4505.

Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M. and Sasaki, M. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry,* 34, 11211–11216.

Thorsen, L., Abdelgadir, W.S., Ronsbo, M.H., Abban, S., Hamad, S.H., Nielsen, D.S. and Jakobsen, M. (2011) Identification and safety evaluation of Bacillus species occurring in high numbers during spontaneous fermentations to produce Gergoush, a traditional Sudanese bread snack. *Int.J.Food Microbiol.,* 146, 244–252.

Treangen, T.J., Abraham, A.L., Touchon, M. and Rocha, E.P. (2009a) Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol.Rev.,* 33, 539–571.

Treangen, T.J., Darling, A.E., Achaz, G., Ragan, M.A., Messeguer, X. and Rocha, E.P. (2009b) A novel heuristic for local multiple alignment of interspersed DNA repeats. *IEEE/ACM Trans.Comput.Biol.Bioinform,* 6, 180–189.

van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M. and Brouns, S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem.Sci.,* 34, 401–407.

Wallace, R.B., Shaffer, J., Murphy, R.F., Bonner, J., Hirose, T. and Itakura, K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res.,* 6, 3543–3557.

Wong, M.L. and Medrano, J.F. (2005) Real-time PCR for mRNA quantitation. *BioTechniques,* 39, 75–85.

Woo, P.C., Lau, S.K., Teng, J.L., Tse, H. and Yuen, K.Y. (2008) Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin.Microbiol.Infect.,* 14, 908–934.

Woo, P.C., Ng, K.H., Lau, S.K., Yip, K.T., Fung, A.M., Leung, K.W., Tam, D.M., Que, T.L. and Yuen, K.Y. (2003) Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. *J.Clin.Microbiol.,* 41, 1996–2001.

Wozniak, R.A. and Waldor, M.K. (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat.Rev.Microbiol.,* 8, 552–563.

Xia, T., SantaLucia, J.,Jr, Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C. and Turner, D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry,* 37, 14719–14735.

Zhang, J. (2003) Evolution by gene duplication: an update. *TRENDS in Ecology and Evolution,* 18, 292–298.

Zhang, Y. and Waterman, M.S. (2005) An Eulerian path approach to local multiple alignment for DNA sequences. *Proc.Natl.Acad.Sci.U.S.A.,* 102, 1285–1290.

Zhao, F., Bai, J., Wu, J., Liu, J., Zhou, M., Xia, S., Wang, S., Yao, X., Yi, H., Lin, M.and et al. (2010) Sequencing and genetic variation of multidrug resistance plasmids in Klebsiella pneumoniae. *PLoS One,* 5, e10141.

Ziemer, C.J. and Steadham, S.R. (2003) Evaluation of the specificity of Salmonella PCR primers using various intestinal bacterial species. *Lett.Appl.Microbiol.,* 37, 463–469.

Zinder, N.D. and Lederberg, J. (1952) Genetic exchange in Salmonella. *J.Bacteriol.,* 64, 679–699.

SUMMARY IN ESTONIAN

PCRi praimeridisaini parendamine

Polümeraasi ahelreaktsioon ehk PCR (ingl. k. Polymerase Chain Reaction) on molekulaarbioloogia valdkonna tehnoloogia, mis võimaldab paljundada spetsiifilist DNA lõiku. DNA paljundamise protsess toimub kolmeosaliste tsüklitena; kõigepealt lahutatakse kaheahelalise DNA järjestuse (*sihtmärkjärjestuse*) ahelad kõrgel temperatuuril, seejärel toimub kahe spetsiifilise DNA järjestuse (*PCRi praimeri*) seondumine sihtmärkjärjestusele praimerite sulamistemperatuuril (T_m) ning seondunud praimerid pikendatakse DNA polümeraasi abil. Kirjeldatud tsüklite kordamine võimaldab PCRi produkti hulga eksponentsiaalset suurenemist. Reaktsiooni käigus paljundatud DNA-d saab detekteerida, kas pikkuse järgi geel-elektorforeesil või reaalajas produkti paljundamise käigus tekkiva signaali abil. PCR võimaldab sel viisil tuvastada erinevatetst DNA proovidest (kliiniline-, veterinaar-, toidu-, keskkonnaproov jne) spetsiifili DNA järjestusi, mistõttu on see tehnoloogia leidnud rakendust erinevates valdkondades. Tänapäeval on järjest enam hakatud kasutama ka DNA järjestamise tehnoloogiat (ehk *sekveneerimist*), mis suudab määrata üksteisele järgnevaid nukleotiide DNA ahelas, kuid paljudes eluliselt olulistes valdkondades kasutatakse valdavalt siiski PCRi tehnoloogiat kindla DNA järjestuse (nt patogeeni) tuvastamiseks proovist.

Paljude erinevate faktorite koosmõju määrab PCRi tundlikkuse ja täpsuse. Täpsus tähendab siinkohal seda, et paljundatakse proovist ainult see DNA järjestus, mida soovitakse tuvastada. Tundlikkus tähendab, et proovist paljundatakse detekteerimiseks piisav kogus soovitud DNA järjestust. Üks olulisemaid eeldusi edukaks PCRi teostamiseks on täpsete ja tundlike praimerite disainimine (*PCRi praimeridisain*). PCRi praimeridisain sisaldab endas erinevaid etappe: sihtmärgi ja sihtmärkjärjestuse valimine (nt kindla järjestuse valimine bakterigenoomist), PCRi praimerijärjestuste disainimine, praimerite täpsuse testimine *in silico* jmt.

Käesolev töö on keskendunud PCRi praimeridisaini erinevate etappide parendamisele; täpsete ja tundlike PCRi praimerite disainimisele ning spetsiifiliste prokarüootsete sihtmärkjärjestuste valimisele ja nende iseloomustamisele. Lühidalt võib käesoleva töö tulemused kokku võtta järgnevalt:

1. Laialt kasutusel oleva vabavaralise PCRi praimeridisaini programmi Primer3 praimerite sulamistemperatuuri (T_m) arvutamise valemi täiustamine (lisades valemisse dupleksi initsiatsiooni entroopiat ja entalpiat arvestava teguri ning sidudes T_m-i arvutamise valemiga kaks erinevat valemit, mis arvestavad igale PCRi reaktsioonile omase soolakontsentratsiooniga ning, millest ühte saab kasutaja, vastavalt enda hinnangule valemi täpsusest, valida) ning sulamistemperatuuri arvutamiseks vajalike termodünaamiliste parameetrite kaasajastamine. Primer3-e eelmise versiooni ja täiustatud versiooni viga praimerite sulamistemperatuuri

arvutamisel (eksperimentaalselt määratud ning arvutuslikult ennustatud väärtuse vahe) on 11.7°C ning 1.37°C, vastavalt.

2. Prokarüootsetele genoomidele suunatud spetsiifilise automaatse praimeridisaini metoodika väljatöötamine ning selle eksperimentaalne valideerimine. Väljatöötatud metoodika tuvastab kõigepealt prokarüootsele liigile unikaalse vähemalt 85 nukleotiidi pikkuse kordusjärjestuse, mis esineb sihtmärkliigi kõigis täissekveneeritud genoomiga tüvedes. Seejärel teostatakse homoloogiaotsingu programmi BLAST abil valitud järjestuse unikaalsuse kontrollimine võrdluses kõigi prokarüootsete liikidega (va sihtmärkliigi) ja inimese genoomi täisjärjestusega. Selliselt leitakse liigispetsiifilised kordusjärjestused. Seejärel disainitakse nimetatud kordusjärjestustele PCRi praimerid programmiga Primer3 ja kontrollitakse (kasutades termodünaamilist lähenemist) disainitud praimerite unikaalsust kõigi prokarüootsete liikide (va sihtmärkliigi) ning inimese genoomi täisjärjestuse vastu. Kirjeldatud metoodika paikapidavust kontrolliti eksperimentaalselt viie mikroobi liigi korral kasutades 132 disainitud PCRi praimeripaari. Statistilise analüüsi käigus (kasutades 132 praimeripaari andmeid) tõestati, et suurema koopiaarvuga kordujärjestuste kasutamine PCRi sihtmärkjärjestustena tagab PCRi suurema tundlikkuse.
3. Eelmises punktis väljatöötatud kordusjärjestusi otsiva metoodikaga leiti 613 kromosomaalse täisjärjestusga liigile spetsiifilised kordusjärjestused. Nimetatud kordusjärjestuste analüüsimisel selgus, et peaaegu kõik prokarüootsed liigid sisaldavad liigispetsiifilisi kordusi, kusjuures 64% liigispetsiifilistest kordustest on seotud valke kodeerivate geenidega ning vaid 14% kordustest on seotud mobiilsete geneetiliste elementidega. Huvitavaks leiuks oli ka, et erinevates prokarüootsetes liikides on funktsioonilt erinev komplekt kordusi, näiteks kui ühes liigis leidub vaid ribosoomi DNA-ga seotud kordusi, siis teises liigis võib leiduda ainult üks suur duplikatsioon, mis omakorda sisaldab arvukalt erinevaid valke kodeerivaid geene.

ACKNOWLEDGEMENTS

First of all, I owe a huge gratitude to my supervisor Professor Maido Remm for providing the possibility to perform this study. For me, his constant motivation and incredible patience are incomprehensible, although encouraging. Thanks to his competence, systematic and calculating thinking, this work has reached its current level.

I thank author of Primer3 software, Steve Rozen, for turning Primer3 to an open source project, and for support in committing the new improvements to SourceForge.

I thank Kai Jõers and Ranno Rätsep from Quattromed HTI Laborid for the experimental work required in Ref II.

I thank Lauris Kaplinski for some help in C.

I thank Tõnu Möls for guidance with the statistics.

I thank all my co-workers in bioinformatics.

Special thanks goes to my lovely course-mate Paula Ann. I thank you for your competent feedback about some scientific questions and organisational issues.

I would like to thank my old course-mate Kadiliina from the bachelor's studies, with whom I passed through the most intensive time during university studies. I appreciate the support you offered.

Also, I like to thank my sister Kadri, who cogitates on whether the science may be stylish. Science™, made in Italy? You have been practically the only member of our family who expressed some support during my PhD *studies*.

PUBLICATIONS

Koressaar T, Jõers K and Remm M (2007).
Enhancements and modifications of primer design program Primer3.
Bioinformatics 23(10): 1289–91

Koressaar T and Remm M (2009).
Automatic identification of species-specific repetitive DNA sequences and
their utilization for detecting microbial organisms.
Bioinformatics 25(11): 1349–55

Koressaar T and Remm M (2012).
Characterization of species-specific repeats in 613 prokaryotic species.
DNAResearch, published online February 24, 2012

CURRICULUM VITAE

I General data

Name	Triinu Kõressaar
Date of birth	27/07/1982, Helme
Citizenship	Estonia
Address	Riia str 23, Tartu, Estonia, 51010
Email	triinu.koressaar@ut.ee
Current position	PhD student, University of Tartu
Education	2004–2006 University of Tartu, MSc
	2000–2004 University of Tartu, BSc

II Scientific activities

My research projects have been focused on PCR primer design, particularly on improving primer design program Primer3 on thermodynamical aspects and designing PCR primers for human pathogens. I have also studied prokaryotic repetitive sequences and included them to PCR primer design.

List of publications:

I. Koressaar T, Jõers K and Remm M (2007). Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10): 1289–91
II. Koressaar T and Remm M (2009). Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms. Bioinformatics 25(11): 1349–55
III. Koressaar T and Remm M (2012). Characterization of species-specific repeats in 613 prokaryotic species. DNAResearch, published online February 24, 2012
IV. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M and Rozen S (2012). Primer3 – New Capabilities and Interfaces. Nucleic Acids Res, published online XXX, 2012

ELULOOKIRJELDUS

I Üldised andmed

Nimi	Triinu Kõressaar
Sünniaeg, koht	27/07/1982, Helme
Kodakondsus	Eesti
Aadress	Riia 23, Tartu, Eesti, 51010
Email	triinu.koressaar@ut.ee
Praegune töökoht, amet	Tartu Ülikool, doktorant
Haridus	2004–2006 Tartu Ülikool, MSc
	2000–2004 Tartu Ülikool, BSc

II Teaduslik tegevus

Minu uurimusprojektid on peamiselt keskendunud PCRi praimeridisainile, täpsemalt täiustades PCRi praimeridisaini programmi Primer3-e termodünaamikaga seonduvaid arvutusi ning disainides praimereid inimese patogeenidele. Samuti olen ma uurinud prokarüootseid korduvaid järjestusi ning kasutanud neid PCRi praimeridisainis sihtmärkjärjestustena.

Publikatsioonide loetelu:

I. Koressaar T, Jõers K and Remm M (2007). Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10): 1289–91
II. Koressaar T and Remm M (2009). Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms. Bioinformatics 25(11): 1349–55
III. Koressaar T and Remm M (2012). Characterization of species-specific repeats in 613 prokaryotic species. DNAResearch, publitseeritud internetis 24ndal veebruaril, 2012a