BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 11 2009, pages 1349–1355
doi:10.1093/bioinformatics/btp241

Genome analysis

Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms

Triinu Koressaar[1], Kai Jõers[2] and Maido Remm[1,*]

[1]Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Riia str. 23, Tartu 51010 and [2]Quattromed HTI® Laborid, Nooruse 9, Tartu 50411, Estonia

Received on January 16, 2009; revised on March 16, 2009; accepted on April 03, 2009
Advance Access publication April 8, 2009
Associate Editor: Alex Bateman

ABSTRACT

Motivation: The concentration of pathogen DNA in biological samples is often very low. Therefore, the sensitivity of diagnostic tests is always a critical factor.

Results: We have developed a novel computational method that identifies species-specific repeats from microbial organisms and automatically designs species-specific PCR primers for these repeats. We tested the methodology on 30 randomly chosen microbial species and we demonstrate that species-specific repeats longer than 300 bp exist in all these genomes. We also used our methodology to design species-specific PCR primers for 86 repeats from five medically relevant microbial species. These PCR primers were tested experimentally. We demonstrate that using species-specific repeats as a PCR template region can increase the sensitivity of PCR in diagnostic tests.

Availability and Implementation: A web version of the method called MultiMPrimer3 was implemented and is freely available at http://bioinfo.ut.ee/multimprimer3/.

Contact: maido.remm@ut.ee

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

1 INTRODUCTION

The concentration of pathogen DNA in biological samples is frequently very low, close to the detection limit. Diagnostic tests based on conventional or real-time PCR often lack the sensitivity required to detect low concentrations of pathogen DNA. The two-step nested-PCR method has occasionally been used to increase the sensitivity (Lopez *et al*., 2003), but this method is time-consuming, more expensive and less reliable than simple one-step PCR.

One way to improve sensitivity is to find an appropriate template region for designing primers (Borst *et al*., 2004). Generally, two types of DNA sequences have been used as template regions for diagnostic PCR: cryptic plasmids and 16S rRNA genes. However, the former cannot be used for all bacteria because many do not contain species-specific plasmids; and in some species, not all strains contain plasmids (Totten *et al*., 1983). It has been reported that ∼5% of *Neisseria gonorrhoeae* strains do not carry the cryptic plasmid that is used to detect this species (Boel *et al*., 2005, Bruisten *et al*., 2004). Therefore, the species cannot be detected in a number of *N.gonorrhoeae*-infected samples or clinical tests using primers designed on the cryptic plasmid. Another problem with targeting plasmid DNA arises from the specificity of the PCR. For instance, it has been found that the sequence of the cryptic plasmid in *N.gonorrhoeae* is highly similar to plasmid sequences in *Neisseria meningitidis* and *Neisseria lactamica* (Ison *et al*., 1986). Because of the high risk that plasmid-based PCR tests will give false-positive and/or false-negative results, many clinical laboratories use confirmatory tests with other PCR primer pairs (Farrell, 1999). Another DNA region typically used as template for diagnostic PCR is the rRNA gene region; rRNA genes are present in the genomes of all microbial species. In fact, they are often present in multiple copies in the genomic DNA, which may increase the sensitivity of their detection (Greisen *et al*., 1994). For example, according to genomic-sequence data, there are 15 copies of rRNA genes in *Photobacterium profundum SS9.* Although the 16S rRNA genes are present as multiple copies in many genomes, it is not always true. There is only one copy of rRNA gene in *Mycobacterium tuberculosis H37Rv.* Additionally, the variability of rRNA gene sequences is not optimal for some tests. For example, 16S rRNA sequences might be too conserved to distinguish closely related species or even strains within the same species. On the contrary, some 16S rRNA sequences may not be conserved enough to design PCR primers with requested parameters. Thus, although the rRNA genes offer a reasonable target region for species detection there is a need for a more systematic approach for selecting target regions.

Repetitive DNA sequences have seldom been used for detecting microbial species because their stability and specificity have been poorly studied. Although repetitive sequences have been described in several microbial species (Achaz *et al*., 2001, 2002, 2003, Skovgaard *et al*., 2002, Ussery *et al*., 2004), they are mostly used to characterize the evolution of bacterial genomes, not for detection or diagnostic purposes. There are only few examples of the use of species-specific repeats to identify pathogens. For example, a 20-copy repeat has been used for *Coxiella burnetii* (Fournier and Raoult, 2003; Klee *et al*., 2006), a 10-copy repeat for *Mycoplasma pneumoniae* (Waring *et al*., 2001), a 7-copy repeat for *Tropheryma whipplei* (Fenollar *et al*., 2003) and a 2-copy repeat for *Brucella spp.* (Leal-Klevezas *et al*., 1995). PCR primers on repetitive sequences have so far been designed manually (Atkins *et al*., 2004, Baker *et al*., 2003).

Species-specific repeats distributed over the genome can offer universal template sequences for diagnostic PCR. Most importantly, the repeat regions could allow microbes to be identified with greater sensitivity. We have developed an automatic method for finding

*To whom correspondence should be addressed.

species-specific repetitive sequences coupled with an automatic procedure for species-specific primer design on those repeats. This approach does not depend on the existence of genome annotations; only the genomic DNA sequences of a target species and non-target species are required. Moreover, specific repeats can be identified not only for species, but also for other clades such as strains, genera or families of micro-organisms.

2 SYSTEM AND METHODS

2.1 Microbial strains

To identify species-specific repeats and to analyze their positions in relation to genes, genomic DNA sequences and gene coordinates were retrieved from the FTP sites of the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/) and NCBI GenBank (ftp.ncbi.nih.gov). The human genome (version NCBI_35) was downloaded from Ensembl (ftp://ensembl.org/pub/). Five microbial genomes were selected for experimental investigation: *N.gonorrhoeae* (*N.gonorrhoeae FA 1090*), *Helicobacter pylori* (*H.pylori 26695* and *H.pylori J99*), *Mycoplasma genitalium* (*M.genitalium G37*), *Listeria monocytogenes* (*L.monocytogenes EGD* and *L.monocytogenes str. 4b F2365*) and *Candida albicans* (*C.albicans SC5314* and *C.albicans WO-1*). Genomic DNA from these strains was ordered from ATCC (http://www.atcc.org/) (*M.genitalium* ATCC 33530D, *N.gonorrhoeae* 700825D, *H.pylori* ATCC 700392D, *L.monocytogenes* ATCC BAA-679 and *C.albicans* ATCC MYA-2876D). Human genomic DNA was obtained from HEK293 (T-rex) cells.

2.2 Algorithm for identifying species-specific repetitive DNA

First, the genomic sequence of the target genome is segmented into splits. The length of a split N and the length of overlap between consecutive splits M must be defined by the user. M is set to 50 bp by default. Matches of sequence splits in the target genome are identified using the similarity search software BLAST. A sequence, split is classified as a candidate repeat if the following three criteria are met: the BLAST match has a bitscore $>1.5*N$; the length of the BLAST match is at least N-15; and the identity between the matching region and the split is >90%. A sequence is defined as a non-species-specific repeat if the length of a BLAST match in any non-target genome has a bitscore $>1.2*N$ and the identity between the match of background sequence and the split is >80%. All scripts are written in Perl programming language.

2.3 PCR primer design algorithm

PCR primers are designed using the program Primer3, version 1.1 (Rozen and Skaletsky, 1998, Koressaar and Remm, 2007). The ability of a primer to amplify alternative PCR products is checked against all available non-target microbial species and the human genome using the program FastaGrep http://bioinfo.ut.ee/download). The following parameters of FastaGrep are used: *-ck 20* (mM concentration of monovalent cations), *-cmg 2.5* (mM concentration of divalent cations), *-cntp 0.8* (mM concentration of NTPs), *-cdna 400* (nM concentration of DNA). The parameter *-dgregions* is used with variable value corresponding to the ΔG of last 12 nucleotides of a given primer. Thermodynamic parameters from the SantaLucia workgroup (SantaLucia and Hicks, 2004) are used to calculate ΔG. Alternative products of a primer pair are computed from the potential binding sites detected by FastaGrep. The maximum length of the alternative PCR products must be defined by the user.

2.4 Experimental testing of PCR primers

In designing the primers for the experiments described in the Section 3, we used some additional parameters that are otherwise set by users. As we did not know which repeat lengths would yield the best-performing primers, we designed primers over a range of split lengths N. Five different lengths were used: 200, 300, 400, 500 and 600 bp. The overlap length M was kept at the default value, 50 bp. The Primer3 flag default values were changed as follows: the primer length range was set to 18–28 bp with optimum at 22 bp, the range of primer melting temperatures was set to 54–63°C with optimum at 57°C, and the range of GC% in the primer sequence was set to 20–80 with optimum at 35%. The maximum length of the predicted alternative product was set to 1000 bp.

All reactions were performed with Biometra T3000 and Eppendorf Mastercycler Gradient PCR machines. The reaction volume was 50 µl, and each reaction mixture contained 10 × PCR buffer B (Solis BioDyne), 5 µl BSA (1 mg/ml, Fermentas), 2 mM of each deoxynucleoside triphosphate, 2.5 mM $MgCl_2$ (Solis BioDyne) and 2 µM of both PCR primers. We used 5 U per µl Taq Hot FIREPol polymerase (Solis BioDyne). The concentrations of genomic DNA template were 10^{-6} (*N.gonorrhoeae*, *L.monocytogenes*, *C.albicans*, and *H.pylori*), 10^{-8} (*M.genitalium*) and 10^{-3} (*Homo sapiens*) µg/µl. All reactions were conducted in the presence of human DNA. All amplifications began with an initial denaturation step at 95°C for 15 min; this was followed by cycles consisting of 1 min at 95°C, 30 s at primer annealing temperature (derived from the melting temperature) and 40 s at 72°C. The number of cycles used in the experiments was 40. PCR products were detected by agarose gel electrophoresis. The lengths and intensities of all bands were registered. Band intensities were assessed visually using a scale from 0 to 5, where 0 indicates a product that was not visually detectable and 5 indicates the most intense band. All experiments were repeated three times and all three results were used separately in statistical analysis.

2.5 Statistical analyses

Statistical analyses were conducted using the software package SAS (SAS Institute Inc. 2004. SAS/STAT User's Guide, Version 9.1.3 SAS Institute Inc., Cary, NC, Vol. 1, Chapter 47, pp. 2659–2851.). To study the intensity of PCR, we used the procedure MIXED, based on the principle of mixed linear models. To evaluate the correlation between PCR intensity and the number of copies per repeat, the SAS procedure CORR was used. More precisely, the Spearman's Rank correlation coefficient was calculated by this procedure.

3 RESULTS

3.1 Computational method to identify species-specific repetitive DNA for PCR primer design

First, we describe the method that we use for automatic identification of species-specific repeated DNA regions in microbial genomes. The prerequisite for finding such regions is the availability of complete genome sequences for the species or strain(s) of interest, called the *target genome* in this article. If more than one genome is selected for identification, then we search for repeats that are present in all group members. In this case, we refer to the *target genome group*. The *non-target genome* is defined as genomic DNA sequence(s) of other organisms potentially present in the sample studied. In this study, the human genome was always used by default as one of the non-target genomes.

A flow diagram of the method for finding specific repeats is shown in Figure 1. The basic steps of the method are as follows. First, genomic DNA sequences of target and non-target species are extracted from the databank. Next, the target genome is split into N bp fragments, which have M bp overlaps with each other. Splits are used as query sequences for a BLAST similarity search to find all similar sequences from the target genome. If two or more matches are found for the given split from the target genome then that sequence split is defined as a candidate species-specific repeat.

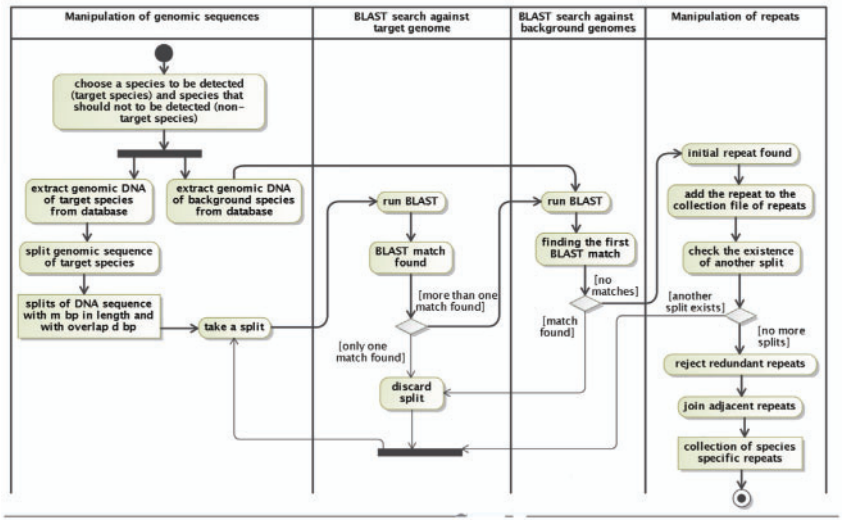

Fig. 1. Action diagram of the process of finding species-specific repetitive DNA sequences in microbial genomes. The whole process can be divided into four sections: procedures with target-genome sequences; finding copies of repeats from the target genome; checking the existence of a repeat in non-target species; and finally, merging and filtering the candidate repeats into final repeat entries. The shortest path to finding a repetitive DNA sequence directly is shown by bold arrows.

All candidate sequence splits are also searched against non-target microbial genomes and against the human genomic sequence. If at least one match is found from any of the non-target genomes then the sequence split is rejected. Redundant splits—two or more splits that are copies of the same repeat—are eliminated. Finally, overlapping splits are joined together to form a final set of target-genome-specific repeats. Detailed parameters of these procedures are described in Section 2.

One of the advantages of our method is that it allows us to find repeats that are specific to more than one genome, i.e. specific to a group of species. In this case, the algorithm randomly chooses one genome from the *target genome group* and searches for species-specific repeats in that genome as described above. The presence of those repeats in other genomes of the *target genome group* is then verified by a homology search against all genomes in that group. If the number of repeat copies in any other genome is <2 then this repeat is discarded. An example of a species-specific repeat from the *N.gonorrhoeae* genome (strains *N.gonorrhoeae FA 1090* and *N.gonorrhoeae NCCP11945*) and its consensus sequence are shown in Figure 2. This repetitive sequence is present in seven copies in both strains analyzed.

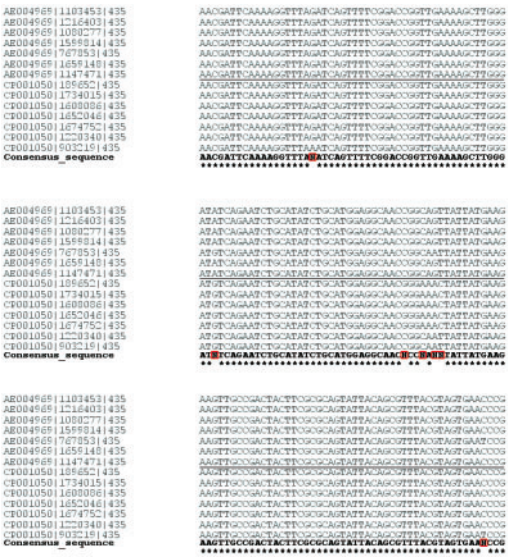

Fig. 2. Flow diagram of the method for designing species-specific PCR primers. The process can be seen as comprising three major steps: making the consensus sequence from copies of the species-specific repeat; designing PCR primers on the consensus sequence; and checking the alternative products from non-target organisms.

3.2 Abundance of species-specific repeats in genomes

To describe the abundance of species-specific repetitive sequences in different microbial genomes (archaea, fungi and bacteria) we randomly picked 30 microbial species out of 508 completely sequenced genomes and searched for repetitive sequences longer than 100, 300 and 1000 bp specific to each species. The numbers of species-specific repeats detected and the maximum number of copies per repeat are shown in Table 1.

Some species in Table 1 have more than one sequenced strain; in these cases, the species-specific repeat identified represents all such strains. Our analyses show that all the 30 genomes studied contain species-specific repeats at least 300 bp long. This length should be sufficient for designing specific PCR primers on the repeat region.

We also tested computationally whether we could find strain-specific repeats for the fully sequenced strains of *Streptococcus pyogenes*, *H.pylori* and *L.monocytogenes.* Such repeats were identified in all three fully sequenced *H.pylori* strains and in both the fully sequenced strains of *L.monocytogenes*. However, 5 of the 12 fully sequenced genomes of *S.pyogenes* strains showed no strain-specific repeats over 100 bp long. This may be explained by the

Table 1. Number of species-specific repeats detected in 30 randomly selected microbial genomes

Species	Number of strains	Number of species-specific repeats			Maximum number of copies per repeat		
		> 100 bp	> 300 bp	> 1000 bp	> 100 bp	> 300 bp	> 1000 bp
Anaplasma phagocytophilum	1	159	87	23	50	23	5
Archaeoglobus fulgidus	1	53	19	4	5	5	5
Arthrobacter aurescens	1	74	26	8	8	6	6
Bdellovibrio bacteriovorus	1	13	8	1	2	2	2
Burkholderia sp.	1	137	17	12	7	5	6
Candidaglabrata CBS138	1	304	132	43	36	25	10
Chlamydophila caviae	1	9	2	1	3	2	2
Chlorobium phaeobacteroides	1	131	61	25	21	21	21
Chlorobium tepidum	1	42	16	4	10	4	3
Cyanobacteria bacterium Yellowstone A-Prime	1	78	24	13	19	16	20
Deinococcus geothermalis	1	19	5	3	5	4	3
Francisella tularensis	6	13	8	1	42	143	3
Haemophilus ducreyi	1	26	13	7	6	6	6
Herminiimonas arsenicoxydan	1	51	13	5	6	5	5
Lactobacillus delbrueckii bulgaricus	2	55	27	10	15	24	14
Magnetospirillum magneticum	1	197	57	27	10	7	10
Mycoplasma gallisepticum	1	74	39	14	8	8	8
Nitrobacter hamburgensis	1	233	81	25	8	23	21
Pseudomonas fluorescens	2	10	5	1	5	5	5
Schizosaccharomyces pombe	1	283	134	54	52	52	14
Staphylococcus aureus	9	24	3	1	5	3	2
S.pyogenes	2	15	6	2	6	71	6
Syntrophobacter fumaroxidans	1	74	30	14	50	13	5
Vibrio vulnificus	2	96	46	11	9	9	9
Zymomonas mobilis	1	34	12	5	5	5	3
L.monocytogenes	2	3	2	0	2	2	0
C.albicans	2	402	71	35	9	17	4
M.genitalium	1	23	5	0	4	2	0
N.gonorrhoeae	1	87	32	16	17	7	7
H.pylori	3	25	3	1	5	12	2

close interrelatedness of some strains (Beres *et al*., 2006, McShan *et al*., 2008), which excludes the possibility of finding strain-specific repeats. In this case, repeats can probably be used to detect groups of closely related strains, but not individual ones.

3.3 Computational design of species-specific PCR primers

After the species-specific repeats have been identified, another part of our computer program automatically designs species-specific PCR primers on those sequences (Fig. 3). In the first step, copies of a repetitive sequence are aligned using ClustalW (Thompson *et al*., 1994). The consensus sequence of all repeats is obtained from the aligned copies by masking non-canonical nucleotides, variable nucleotides and locations of insertions/deletions with the symbol N in the first sequence in the alignment. The use of N in the consensus sequence precludes the design of primers at ambiguous nucleotides; by default, Primer3 rejects primer candidates containing the symbol N. In the second step, PCR primers on the consensus sequence are designed using the primer design program Primer3, version 1.1 (Koressaar and Remm, 2007; Rozen and Skaletsky, 1998). In the third step, the potential binding sites of candidate primers are identified in non-target microbial species and human genomic DNA. The binding sites in non-target genomes are searched using the program FastaGrep (http://bioinfo.ut.ee/download/). Finally, alternative products from non-target genomes are calculated for all candidate primer pairs. In the fourth step, the primer pairs that could potentially yield alternative products from non-target genomes are discarded.

3.4 Experimental testing of primers for repeat regions

To test whether the design of primers on species-specific repeats increases the sensitivity of PCR in practice, we chose five microbial genomes (*N.gonorrhoeae*, *M.genitalium*, *L.monocytogenes*, *H.pylori* and *C.albicans*) and designed such primers. Our objective was to design 10 primer pairs per repeat and to test whether the signal intensity is correlated with the number of copies per repeat. We designed 1956 different candidate primer pairs for 193 different repetitive DNA sequences, an average of about 10 primer pairs per repeat. For experimental tests, we selected 132 different primer pairs designed on repeats with 2–16 copies and other pairs amplifying non-repeated regions (as we are interested in sensitivity of PCR when designing this experiment, for simplicity we omitted the third and fourth primer design step here).

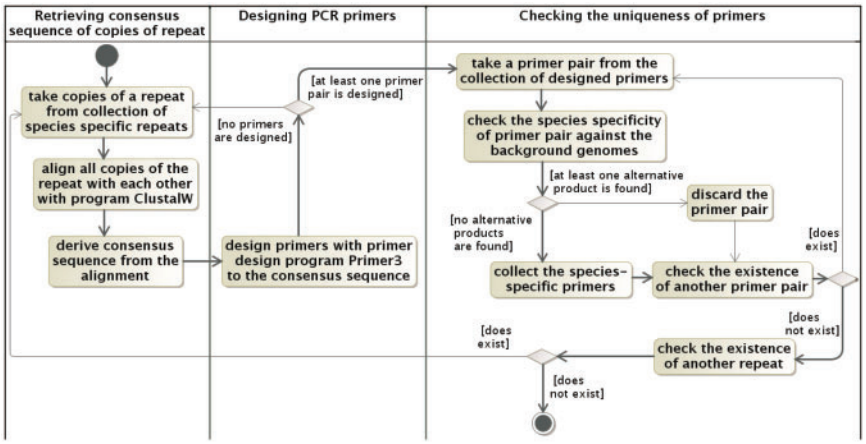

Fig. 3. Action diagram of the method for designing species-specific PCR primers. The process can be seen as comprising three major steps: making the consensus sequence from copies of the species-specific repeat; designing PCR primers on the consensus sequence; and checking the alternative products from non-target organisms.

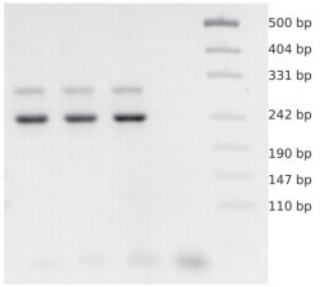

Fig. 4. The intensity of the PCR product band is dependent on number of copies. Intensities of gel electrophoresis bands with different numbers of copies in the *C.albicans* genome. The faint upper band (298 bp) identifies a PCR product amplified from a single target in the genome. The brighter lower band (238 bp) identifies a multi-target PCR product with 10 copies of repeats in the same genome. Both these target regions contain 100% identical binding sites for the primers.

We assessed sensitivity visually by assigning gel electrophoresis bands the intensity values in the range between 1 and 5. An example of gel electrophoresis of two PCR products with different lengths, each with different repeat copy numbers in the genome, is illustrated in Figure 4.

We tested whether there is a statistical correlation between the number of copies per repeat and the gel electrophoresis band intensity of PCR primers designed onto these repeats. This relationship was examined by regression analysis using the mixed model. It was statistically significant: the F-statistic (20.17, DF = 1) for the linear term of the factor of the number of copies per repeat gives $P < 0.0001$. The positive correlation between the number of copies per repeat and the gel electrophoresis band intensity was also verified by Spearman's Rank correlation coefficient. The correlation coefficient was 0.38 ($P < 0.0001$).

These results indicate that PCR using species-specific repeats as targets can be more sensitive than PCR targeting single-copy sequences in the same genome. The PCR product lengths in our experiments were variable (150–650 bp). As longer PCR products can appear stronger and easier to detect, it was important to verify that the PCR product lengths were not biased towards high-copy repeats. We observed no correlation in our test set between the copy number of repeats and PCR product length confirming that the increase in sensitivity is not caused by unequal distribution of PCR product lengths. Figure 5 is a visual illustration how the fraction of experiments with high-intensity gel electrophoresis band increases with increasing copy numbers of the repeat, particularly for repeats with more than four copies. The figure shows results for primers designed onto the repeats having 1–8 number of copies. There were too little primer pairs to reliably estimate the intensity of primer pairs designed onto the repeats with copy number >8. Despite the clear and statistically significant trend, there are fluctuations. We have no reason to think that they are technical artifacts. They are probably caused by the limited number of tested repeats and/or species. The low intensity of repeats with seven copies can be influenced by the fact that repeats with seven copies were tested in one specific genome only. However, when calculating the F-statistic above, the influence caused by the different species (it was significant $P < 0.01$) is excluded. Thus, we assume that having 5–8 copies of repeat yields optimal intensity of the gel electrophoresis band in diagnostic PCR.

3.5 Web interface

The software implementing our algorithm is called MultiMPrimer3 and is located at http://bioinfo.ut.ee/multimprimer3/. Currently, users can design diagnostic PCR primers by choosing the target species of interest (or strains of a given species) from a predefined list of bacterial species and strains. Alternatively, a user can insert his/her own sequences or file containing sequences in FASTA format to yield PCR primers specific for those sequences.

3.6 Performance analysis

Runtime is mainly limited by time spent on finding repetitive sequences (which is in turn affected by the number of chosen target genomes, the homology/similarity of target genomes with each other and with non-target genomes, the length of target/non-target genome sequences, e.g human genome). For instance, choosing four strains

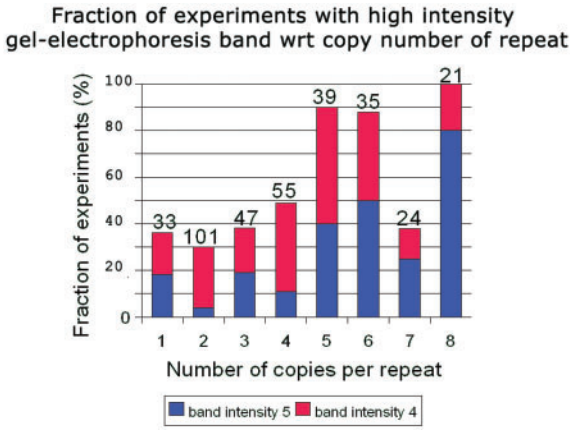

Fig. 5. The fraction of PCR experiments for which gel electrophoresis band intensity is 4 (indicated by light gray column) or 5 (indicated by dark gray column) according to the number of copies per repeat. The scale for assessment of band intensities ranges from 0 to 5. Thus, the intensities 4 and 5 represent the most intensive bands. The number of PCR experiments carried out for each type of repeat is shown on stop of each column.

of *Chlamydophila pneumoniae* as target genomes and three genomes from the same genera but different species as non-target genomes takes about 5 min. However, when choosing more complex genomes as target genomes (group of genomes that are not so closely related or group of species having closely related non-target species) or human as a non-target genome, the runtime may exceed 24 h.

4 CONCLUSION

We have developed a new approach to PCR-based detection of microorganisms. It utilizes genomic sequences to find species-specific repeats and designs PCR primers on those repeats. As we show, such species- or group-specific repeats can increase the sensitivity of detection. One advantage of our method is that it allows different phylogenetic groups to be handled automatically—microbial genera, group of species or group of strains. Thus, the method introduced here offers a universal approach to designing taxon-specific PCR primers. Another advantage is that annotations of target and non-target organism genomes are not required, in contrast to certain other approaches (Fredslund *et al*., 2005, 2007; Fu *et al*., 2008; Kim and Lee, 2007). The taxon-specific repetitive sequences are searched automatically from genomic sequences and the algorithm does not require knowledge of the location of rRNA or other genes. However, in case of genomic sequences in the assembly stage, alternative method for finding species-specific repeats should be considered. For example, using method that finds repeats from Genome Survey Sequences dataset (Otto *et al*., 2008).

The previously proposed idea that long repeats in bacterial genomes are common to all prokaryotes (Rocha *et al*., 1999) was well supported by our findings. Moreover, species-specific repeats were found in all the microbes analyzed (Table 1). The stability of repeat sequences is an important issue; repeats are not usable for diagnostic purposes if they can disappear in few generations. This issue remains poorly studied. Interspersed repeats are commonly deleted by homologous recombination and frequently with the loss of genetic material. Thus, interspersed repeats persist more readily than tandem repeats (Achaz *et al*., 2002). Furthermore, some repeats that persist in bacterial genomes may have a selective advantage for some species (Achaz *et al*., 2003). Thus, usage of repeats as PCR template regions for identifying microbes should be possible for wide range of microbes. Our methodology can detect both interspersed and tandem repeats. It is posssible to select one or another type of repeats for primer design if user prefers to do so.

The definition of a species-specific repeat is sophisticated and is directly dependent on the detection procedure. There will always be some similarity between species-specific repeat and non-target genome sequences. Therefore, it is important to realize that the first step of our algorithm, detection of species-specific repeats, is not sufficient to guarantee specific detection of species using those repeats. Rather, the repeats serve as candidates for PCR primer design. It is the PCR primer design and validation procedure that ensures specific detection of a given species. The primer design procedure is crucial for distinguishing one species from others. In step two of our algorithm, we search for primer binding sites in non-target genomes. Most existing primer design software packages use sequence similarity searches to identify non-specific PCR products. Our algorithm models the primer binding sites using a thermodynamic model of DNA–DNA binding, which allows for more correct prediction of PCR primer binding and subsequent synthesis of detectable non-specific PCR products (Andreson *et al*., 2008). We consider thermodynamic modeling of PCR primer binding sites using FastaGrep software to be one of the main advantages of our methodology.

We performed our experiments using conventional PCR, whereas many diagnostic tests use real-time quantitative PCR. However, primer design principles for conventional PCR and real-time quantitative PCR are quite similar. Real-time PCR requires a product of optimal length (Hird *et al*., 2006, Livak and Schmittgen, 2001, Mouillesseaux *et al*., 2003, Park and Crowley, 2005), which can be adjusted during the primer design stage of our algorithm. Thus, the methodology and software we describe here should be suitable for designing both conventional and real-time PCR primers.

ACKNOWLEDGEMENTS

The authors would like to thank Reidar Andreson and Paula Ann Kivistik for critical reading of the manuscript.

Funding: Enterprise Estonia (EU19730); Estonian Science Foundation (grant ETF#6041); Estonian Ministry of Education and Research (grant 0182649s04); EU (through the European Regional Development Fund through the Estonian Centre of Excellence in Genomics).

Conflict of Interest: none declared.

REFERENCES

Achaz,G. *et al*. (2001) Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol*., **18**, 2280–2288.

Achaz,G. *et al*. (2002) Origin and fate of repeats in bacteria. *Nucleic Acids Res*., **30**, 2987–2994.

Achaz,G. *et al*. (2003) Association between inverted repeats and the structural evolution of bacterial genomes. *Genetics*, **164**, 1279–1289.

Andreson,R. *et al*. (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Res*., **36**, e66.

Atkins,S.D. and Clark,I.M. (2004) Fungal molecular diagnostics: a mini review. *J. Appl. Genet*., **45**, 3–15.

Baker,G.C. *et al.* (2003) Review and re-analysis of domain specific 16S primers. *J. Microbiol. Methods*, **55**, 541–555.

Beres,S.B. *et al*. (2006) Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A Streptococcus. *Proc. Natl Acad. Sci. USA*, **103**, 7059–7064.

Boel,C.H.E. *et al*. (2005) Evaluation of conventional and real-time PCR assays using two targets for confirmation of the COBAS AMPLICOR CT/NG test for Neisseria gonorrhoeae in clinical samples. *J. Clin. Microbiol*., **43**, 2231–2235.

Borst,A. *et al*. (2004) False-positive results and contamination in nucleic acid amplification assays: suggestions for a prevent and destroy strategy. *Eur. J. Microbiol. Infect. Dis.*, **23**, 289–299.

Bruisten,S.M. *et al*. (2004) Multicenter validation of the cppB gene as a PCR target for detection of Neisseria gonorrhoeae. *J. Clin. Microbiol*., **42**, 4332–4334.

Farrell,D.J. (1999) Evaluation of AMPLICOR Neisseria gonorrhoeae PCR using cppB nested PCR and 16S rRNA PCR. *J. Clin. Microb*., **37**, 386–390.

Fenollar,F. *et al*. (2003) Use of genome selected repeated sequences increases the sensitivity of PCR detection of Tropheryma whipplei. *J. Clin. Microbiol.*, **42**, 401–403.

Fournier,P.E. and Raoult,D. (2003) Comparision of PCR and serological assays for early diagnosis of acute Q fever. *J. Clin. Microbiol.*, **41**, 5094–5098.

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

Fredslund,J. *et al*. (2005) PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucleic Acids Res*., **33**, W516–W520.

Fu,Q. *et al*. (2008) PRISE (PRImer SElector): software for designing sequence-selective PCR primers. *J. Microbiol. Methods*, **72**, 263–267.

Greisen,M.L. and Purohit,A. (1994) PCR primers and probes for the 16S rRNA gene of most species of pathogenic bacteria, including bacteria found in cerebrospinal fluid. *J. Clin. Microbiol.*, **32**, 335–351.

Hird,H. *et al*. (2006) Effect of heat and pressure processing on DNA fragmentation and implications for the detection of meat using a real-time polymerase chain reaction. *Food Addit. Contam.*, **7**, 645–650.

Ison,C.A. *et al*. (1986) Homology of cryptic plasmid of Neisseria gonorrhoeae with plasmids from Neisseria meningitidis and Neisseria lactamica. *J. Clin. Pathol*., **39**, 1119–1123

Kim,N. and Lee,C. (2007) QPRIMER: a quick web-based application for designing conserved PCR primers from multigenome alignments. *Bioinformatics*, **23**, 2331–2333.

Klee,S.K. *et al*. (2006) Highly sensitive real-time PCR for specific detection and quantification of Coxiella burnetii. *BMC Microbiol*., **6**, 2.

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

Leal-Klevezas,D.S. *et al*. (1995) Single-step PCR for detection of Brucella spp. from blood and milk of infected animals. *J. Clin. Microbiol.*, **33**, 3087–3090.

Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCt method. *Methods*, **25**, 402–408.

Lopez,M.M. *et al*. (2003) Innovative tools for detection of plant pathogenic viruses and bacteria. *Int. Microbiol.*, **6**, 233–243.

McShan,W.M. *et al*. (2008) Genome sequence of a nephritogenic and highly transformable M49 strain of streptococcus pyogenes. *J. Bacteriol.* [Epub ahead of print, doi:10.1128/JB.00672-08, September 26, 2008].

Mouillesseaux,K.P. *et al*. (2003) Improvement in the specificity and sensitivity of detection for the Taura syndrome virus and yellow head virus of penaeid shrimp by increasing the amplicon size in SYBR Green real-time RT-PCR. *J. Virol. Methods*, **111**, 121–127.

Otto,T.D. *et al*. (2008) ReRep: computational detection of repetitive sequences in genome survey sequences (GSS). *BMC Bioinformatics*, **9**, 366.

Park,J.W. and Crowley,D.E. (2005) Normalization of soil DNA extraction for accurate quantification of target genes by real-time PCR and DGGE. *BioTechniques*, **38**, 579–586.

Rocha,E.P.C. *et al*. (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in bacillus subtilis and other competent prokaryotes. *Mol. Biol. Evol.*, **9**, 1219–1230.

Rozen,S. and Skaletsky,H. (1998) Primer3. Available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html (last accessed date 5 January, 2009)

SantaLucia,J. Jr and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.

Skovgaard,M. *et al*. (2002) The atlas visualisation of genome-wide information. *Meth. Microbiol.*, **33**, 49–63.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673–4680.

Totten,P.A. *et al*. (1983) DNA hybridization technique for the detection of Neisseria gonorrhoeae in men with urethritis. *J. Infect. Dis.*, **148**, 462–471.

Ussery,D.W. *et al*. 2004 Genome update: DNA repeats in bacterial genomes. *Microbiology*, **150**, 3519–3521.

Waring,A.L. *et al*. (2001) Development of genomics-based PCR assay for detection of Mycoplasma pneumoniae in a large outbreak in New York state. *J. Clin. Microbiol.*, **39**, 1385–1390.