Genome-scale design of PCR primers and long oligomers for DNA microarrays

Stefan A. Haas*, Marc Hild1, Anthony P. H. Wright2,3, Torsten Hain4, Driss Talibi5 and Martin Vingron

Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Ihnestraße 73, D-14195 Berlin, Germany, 1Center for Molecular Biology Heidelberg (ZMBH), University of Heidelberg, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany, 2Natural Sciences Section, Södertörns University College, S-141 89 Huddinge, Sweden, 3Department of Biosciences, Karolinska Institute, Novum, S-141 57, Stockholm, Sweden, 4Department of Medical Microbiology, Justus-Liebig University School of Medicine, Frankfurter Straße 107, 35392 Giessen, Germany and 5Eurogentec SA, Parc Scientifique du Sart Tilman, 4102 Seraing, Belgium

Received June 24, 2003; Revised and Accepted August 4, 2003

ABSTRACT

During the last years, the demand for custom-made cDNA chips/arrays as well as whole genome chips is increasing rapidly. The efficient selection of gene-specific primers/oligomers is of the utmost importance for the successful production of such chips. We developed GenomePRIDE, a highly flexible and scalable software for designing primers/oligomers for large-scale projects. The program is able to generate either long oligomers (40–70 bases), or PCR primers for the amplification of gene-specific DNA fragments of user-defined length. Additionally, primers can be designed in-frame in order to facilitate large-scale cloning into expression vectors. Furthermore, GenomePRIDE can be adapted to specific applications such as the generation of genomic amplicon arrays or the design of fragments specific for alternative splice isoforms. We tested the performance of GenomePRIDE on the entire genomes of Listeria monocytogenes (1584 gene-specific PCRs, 48 long oligomers) as well as of eukaryotes such as Schizosaccharomyces pombe (5006 gene-specific PCRs), and Drosophila melanogaster (21306 gene-specific PCRs). With its computing speed of 1000 primer pairs per hour and a PCR amplification success of 99%, GenomePRIDE represents an extremely cost- and time-effective program.

INTRODUCTION

DNA microarrays are currently the method of choice for the large-scale analysis of gene expression. These arrays are mainly designed for the investigation of entire transcriptomes aiming to detect new genes or to unravel metabolic pathways (1–3). In pharmacology, custom-made chips carrying only a few hundred candidate genes are frequently used for diagnostics of complex diseases (4–7), for the comparison of different cell types (8) or developmental stages (9).

The design strategy for DNA microarrays depends strongly on the precise application planned for each particular array (10) but is also limited by the costs and local resources (technical equipment, access to clone libraries, etc.). cDNA microarrays are frequently used for the analysis of tissue-specific gene expression and/or the comparison of tumor and normal tissues (4,6). The quality of clone-based microarrays may suffer from contamination of clones by foreign DNA, and also from the inconsistent size range of the spotted clones, thus complicating quantitative evaluation of hybridization signals. Furthermore, since the spotted fragments are not designed, there can be problems of cross-hybridization with non-cognate sequences in the probe mixture.

The potential disadvantages of cDNA microarrays are avoided in the design of PCR/oligo arrays. The knowledge of genome sequence and its annotation of genes can be used to select subsequences that uniquely identify genes. Ideally, such gene-specific sequences will be subject to PCR amplifications that result in fragments of constant length, thus minimizing the differences in PCR amplification efficiency as well as in hybridization kinetics. Alternatively, either a small set of short oligonucleotides (1) or a single long oligomer (11) (50–70 bp) representing a gene can be selected. Kane et al. (12) showed that those long oligomers exhibit a higher sensitivity in the detection of their target gene than conventional short oligomers. However, the risk of cross-hybridization increases compared with the use of short match and mismatch oligomers (13).

Up to now, several primer/oligo design programs have been developed. All of them are specialized for either the design of PCR primers (14–17) or oligomers (18,19). Our GenomePRIDE software provides a more flexible approach of designing primers for many applications, but also outperforms other software in quality and speed.

*To whom correspondence should be addressed. Tel: +49 30 8413 1164; Fax: +49 30 8413 1152; Email: haas@molgen.mpg.de
MATERIALS AND METHODS

PCR amplification

Primers for the Drosophila melanogaster, Schizosaccharomyces pombe and Bacillus subtilis project were synthesized by Eurogentec SA (Belgium). For Listeria monocytogenes primers were ordered from Metabion (Germany). PCR amplifications were performed in 96-well microtiter plates. The Drosophila ORFs were initially amplified from genomic DNA using gene-specific primers, all of which contained one of 10 amplicon unspecific common tag sequences of 15 nt length at their 5’ ends. The use of different tags reduces the risk of cross contamination. Subsequent re-amplification was carried out using the matching tag–primer pairs. The first PCR round was performed in 50 μl reactions containing 100 ng of genomic DNA, 1× Qiagen PCR buffer (1.5 mM MgCl₂), 40 μM of each dNTP, 1 U Qiagen Taq polymerase and 20 pmol of each primer. The plates were incubated for 5 min at 94°C, followed by 35 cycles of denaturation at 94°C for 30 s, annealing for 30 s and elongation at 72°C for 90 s. The annealing temperature was lowered during the first 10 cycles from 65 to 50°C to increase specificity of the amplification. In the last 20 cycles the elongation time was prolonged by 5 s for each cycle to compensate for decreasing enzymatic activity. Re-amplification was initiated by inoculating a 100 μl PCR with 1 μl of first round reaction containing 1× Qiagen PCR buffer (1.5 mM MgCl₂), 0.1 mM of each dNTP, 2 U Qiagen Taq polymerase and 50 pmol of each primer. The plates were incubated for 5 min at 94°C, followed by 35 cycles of denaturation at 94°C for 30 s, annealing for 30 s at 60°C and elongation at 72°C for 90 s. In the last 20 cycles the elongation time was prolonged by 5 s for each cycle to compensate for decreasing enzymatic activity. Amplification success and the length of the resulting PCR fragments for first and second PCRs were checked on 1% agarose gels.

Genomic sequences

The D.melanogaster project was based on release 2 of the Berkley Drosophila Genome Project (BDGP, http://www.fruitfly.org/sequence/download.html) in combination with an independent prediction using the software GENESHE (V. Solovyev). We used the S.pombe genome sequence and annotations downloaded from the Sanger Center (ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Chromosome_contigs, 22 Mar 2002). The L.monocytogenes genome data are based on sequence NC_003210 (21 October 2001).

Thermodynamic stability computation

We used the free energy ΔG as a measure of the thermodynamic stability of a primer-binding site. For primers shorter than 30 bp, ΔG is computed using the following equation [ΔH and ΔS derived from Sugimoto et al. (20)]:

\[ ΔG = ΔH - T * ΔS \]

where ΔH describes the enthalpy and ΔS the entropy. The temperature T equals the user-defined melting temperature. GenomePRIDE aims to keep ΔG of a primer constant, consequently also fixing the melting temperature of the resulting primer/oligo. This linear correlation between the user-defined melting temperature and the resulting melting temperature is no longer valid for long oligomers since the melting temperature of a 70mer would be ~90°C. Therefore, we interpolated the temperature T for oligomers of length 40, 50, 60 and 70 bp by computing T for a random set of oligomers of fixed length while also keeping ΔG fixed to the optimal value used for short oligomers.

Software integration

GenomePRIDE is implemented as a module of the freely available Staden package software (21) (gap4), which is usually used for large-scale sequencing projects. This tight integration enables GenomePRIDE to use many features of the Staden package that help to manage/visualize whole genome PCR projects dealing with large sequences and their annotations. One of these features is the tagging of subsequences according to annotations like coding sequence, exons, repeats, etc.

In general, GenomePRIDE makes use of nine different types of tags (Table 1) to represent annotations relevant to PCR primer design: coding sequence (CODG or PCRT), exon (EXON), repetitive sequence (REPT), and similarity regions between genes (AMBG, AMBI). Optionally, a description can be attached to each individual tag. CODG/EXON tags mark the subsequences that will be the subject of the primer design. Usually, they reflect parts of genes, where the first line of the tag description defines the name of the gene, and the second line contains the number of the respective exon. In this way,
GenomePRIDE is able to analyze all exons of a gene by evaluating all CODG/EXON tags sharing the same gene name. Per definition AMBG/AMBI tags result from a BLAST (22) comparison of every ‘unspliced’ transcript (exons + introns) versus all other transcripts. Regions within the exons of the query that show similarity to exons of other genes are marked as AMBG tags, whereas similarities to introns are represented as AMBI tags. Although the tags may highlight discrete features of a sequence, GenomePRIDE does not check the validity of these annotations in order to keep GenomePRIDE customizable for many applications.

The sequence tags can be uploaded from EMBL file format, such that the feature table entries are translated into different tags. Alternatively, tags might be defined by the user either manually using a graphical editor, or automated by batch-wise generation of a text file of tags (Fig. 1) via self-made scripts.

Tools and configuration

A prerequisite for the design of primers/oligomers by GenomePRIDE is the appropriate assembly of the target genome and annotations into the Staden package. In order to facilitate data import, GenomePRIDE provides tools for the automated upload of projects dealing either with the design of gene-specific PCR fragments or the generation of genomic amplicons. Since whole genome projects may contain thousands of sequence tags, additional tools address the management of these tags. The tool ‘Show Tags’ can be used to localize and visualize discrete subsets of tags thus facilitating manual inspection of the surrounding sequence. In the case of falsely placed tags, ‘Remove Tags’ allows their batch-wise removal.

GenomePRIDE provides a configuration interface in order to adjust several parameters involved in the selection process of the appropriate target sequence for primer/oligo design. These options include the definition of the optimal length of the resulting PCR product, either by fixing its size or by forcing GenomePRIDE to cover the entire gene. Furthermore, the user can also define the preferred relative location of a fragment with respect to the gene such that a fragment partially covering a gene is forced to cover either the 3’ terminus, 5’ terminus or the central part of the gene. In order to target a primer to a specific position of a gene the primer can be fixed either at its 5’ or 3’ terminus. Additional options are described in the manual (http://pride.molgen.mpg.de/genome_pride_toc.html).

**Figure 1.** Example of a text file describing two overlapping sequence tags. The line starting with ‘ID’ defines the contig on which the tag has to be placed. The following line describes the type of the tag, its orientation (+/−) and the start/end positions. Further ‘TC’ lines contain comments like the name of the corresponding gene or some description.

**Table 1.** First six lines of a given tagging file, containing the identifier (ID), type (TC) and comments of sequence tags.

<table>
<thead>
<tr>
<th>ID</th>
<th>chr4</th>
<th>TC</th>
<th>CODG + 1..1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>fragment1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>some comment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>chr4</td>
<td>TC</td>
<td>CODG + 901..1900</td>
</tr>
<tr>
<td>TC</td>
<td>fragment2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>some comment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.** Annotations of sequences interpreted by GenomePRIDE. The Staden package allows the marking of sequence features [exons (orange boxes), repeats (green boxes), homology regions (gray boxes)] by specific tags (Table 1) sketched in (A). Based on these tags, GenomePRIDE generates an artificial subsequence of the genome that includes all sequence information relevant for the primer/oligo design for a single gene (B). This sequence consists of a concatenation of all exons flanked by upstream/downstream sequences (dashed lines) of a user-defined length (default 10 kb). In order to enable GenomePRIDE to check for secondary binding sites within the introns of the gene of interest, these intronic sequences are appended. Similarities to more distant genes are still represented by the appropriate tagging.

Design strategy

The overall strategy of designing gene-specific PCR primers is divided into two phases. Given the annotation of the exon structure of the gene of interest, GenomePRIDE first screens for the optimal target region within those exons. In a preprocessing step, GenomePRIDE splices the genomic sequence covering the target gene in order to generate an artificial sequence consisting of a concatenation of all related exons followed by all spliced intron sequences (Fig. 2). GenomePRIDE now computes a quality for every potential fragment of a user-defined length within the exonic sequence by evaluating the fraction of homology to other genes, the fraction of intronic sequence, and by measuring the location of the putative target region with respect to the preferred location defined by the user. If no target region reaches a quality above a certain threshold, the optimal fragment length is automatically reduced to 50% in order to increase the likelihood of finding a target region of good quality. For example, in cases where all exons are shorter than the user-defined fragment length the fraction of intron sequence would be high, leading to the automatic reduction of optimal fragment length to 50% of the original length. This procedure is repeated as long as no region above the quality cut-off is found, and the reduced length is still longer than the minimal fragment length. However, if none of the target regions reaches the quality threshold, a region of the original optimal length will be selected.

After defining the optimal region within a gene, GenomePRIDE computes both PCR primers independently. The optimal position of a primer is hereby defined by the boundaries of the previously selected target region, aiming to amplify a fragment of optimal length. Similar to the PRIDE software (23) used for sequencing, the design of a single primer using GenomePRIDE is just based on the evaluation of the thermodynamic stability, strength of the most stable secondary binding site, formation of primer dimers and the position of the primer (now with respect to the preselected target). The evaluation of potential secondary binding sites of each primer includes all exons and introns of the respective gene, but also includes the sequences flanking the gene.
(Fig. 2B). The user defines the length of these upstream/downstream regions. By default, the length of these flanking regions is 10 kb. In the case of the mispriming of one of the primers, this is generally sufficient to avoid the amplification of a secondary PCR product. If both primers have secondary binding sites that may give rise to an additional PCR product somewhere else, these sites are usually located in regions that were already detected by the BLAST similarity search.

The selection of the optimal target region for the design of long oligomers is performed in the same way as for PCR primers. The basic parameters in primer design are also used as a measure of the oligomer quality; however, the thermodynamic stability of the 3’ terminal bases is not evaluated.

RESULTS


Table 2 shows the statistics of the PCR primers used for the two largest primer sets for D.melanogaster and S.pombe comprising 21 306 and 5006 primer pairs, respectively. With an overall amplification success of 99%, GenomePRIDE outcompetes other primer design software. A large fraction of the resulting PCR fragments is close to the user-defined optimal fragment length (Fig. 3). Much larger or shorter PCR products are usually caused by genes only consisting of short exons, or by genes that share similarities to other genes. Although GenomePRIDE does not evaluate the GC content of primers explicitly, their average GC content roughly matches the overall GC content of the coding sequences of the respective organisms. The GC content of a small number of primers deviates significantly from the overall GC content (<30% or >70% GC). Nevertheless, these unusual primers that might not be accepted by other primer design programs did not cause a lower success in amplification. According to the variations in GC content, the average length of the PCR primers differs between the organisms. The number of PCR fragments carrying intron sequences or similarity regions was <6% of the total number of fragments, but these numbers strongly depend on the structure of the predicted genes.

With all 48 long oligomers used in the L.monocytogenes project we were able to reproduce the results obtained when using the corresponding gene-specific PCR fragment. These data were also confirmed by northern blots (data not shown).

Figure 3. Length distribution of PCR fragments (D.melanogaster). A large fraction of the PCR fragments matches the user-defined fragment length of 500 bp. GenomePRIDE designed some fragments including intronic sequence up to a total length of ~600 bp since a small amount of intron sequence (<15%) is tolerated. Only a few fragments are longer than 600 bp, this being caused by genes with exons shorter than the minimal fragment length of 100 bp.

<table>
<thead>
<tr>
<th></th>
<th>Drosophila melanogaster</th>
<th>Schizosaccharomyces pombe</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of primer pairs succeeded/failed</td>
<td>21 306/90</td>
<td>5006/33</td>
</tr>
<tr>
<td>Successful amplifications</td>
<td>20 948 (98.4%)</td>
<td>4997 (99.8%)</td>
</tr>
<tr>
<td>No. of fragments including introns (intron &gt;15% of optimal length)</td>
<td>1300</td>
<td>97</td>
</tr>
<tr>
<td>No. of fragments including homology</td>
<td>1496</td>
<td>266</td>
</tr>
<tr>
<td>GC content (primers/ORFs) (%)</td>
<td>51/52</td>
<td>40/40</td>
</tr>
<tr>
<td>Average length of primers (bp)</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>Computation time (h)</td>
<td>23</td>
<td>5</td>
</tr>
</tbody>
</table>
genomic context of the *bithorax* and *antennapedia* complex, respectively. All primer pairs resulted in fragments of appropriate size (M.Hild, Center for Molecular Biology Heidelberg, personal communication).

GenomePRIDE is able to compute ~1000 primer pairs per hour on a standard Linux-PC using 50–800 Mb of main memory. All primers/oligomers generated are stored in a tab-delimited text file listing the name, the sequence and key features.

**DISCUSSION**

In contrast to other PCR primer design software (14–17), GenomePRIDE was tested on large data sets including entire eukaryotic and prokaryotic genomes. Only the GST-PRIME (17) software was tested on a subset of genes of eukaryotes, but yielded a lower record of success than GenomePRIDE. The PRIMEGENS (14) and PrimeArray (15) software [based on primer3 (16)] may not be directly applicable to eukaryotic sequences since the more complex gene structure complicates the selection of optimal primers. While GenomePRIDE was successfully applied to large-scale expression cloning, RNAi and alternative splicing experiments, none of the existing programs supports these applications. GenomePRIDE both computes primers/oligomers of high quality and provides tools for visualization and management of large projects by using standard features of the Staden package. In addition, the analysis of the quality of potential primers is much different than in other software packages since only a few basic parameters (thermodynamic stability, stability of secondary binding sites and dimer formation) are evaluated. Parameters like GC content, loop formation, specific bases at the 3' terminus of primers or the formation of heterodimers of two PCR primers are not analyzed, mainly because most of these parameters are not independent from the parameters used by GenomePRIDE. For instance, primers able to form loops are also able to form primer dimers, thus dimer formation sufficiently reflects both criteria. Additionally, most software aims to design primers with an intermediate GC content. This reflects the fact that primers with extreme GC content are more likely to form dimers or to have a stable secondary binding site. Consequently, the extensive evaluation of dimer formation and the computation of the thermodynamic stability of the most stable secondary binding site leads indirectly to primers with an average GC content. Furthermore, a heterodimer-forming primer will show potential secondary binding at the site of the second primer since the sequence of both primers and their complements are always included in secondary binding analysis. However, this clear focus on the evaluation of secondary binding has the consequence that GenomePRIDE relies on the availability of sequences flanking the genes of interest. The more flanking sequence is provided for this analysis, the more likely it does include unfavorable sequences (e.g. repeats) that GenomePRIDE will subsequently avoid. On the other hand, if GenomePRIDE is forced to check several 100 kb for secondary binding the likelihood of finding stable secondary binding sites will increase even though they might not be relevant in subsequent PCR or hybridization reactions.

The primary focus of GenomePRIDE is the generation of PCR primers for gene-specific PCR based on a given genome sequence and the corresponding annotation of genes. More specifically, the software also supports the design of primers in a specific reading-frame of a coding region (defined by CODG tags), thus facilitating the application of the resulting PCR fragments in expression cloning. In addition to the design of PCR primers, GenomePRIDE also provides a means for the design of long oligonucleotides (40–70 bp).

Moreover, GenomePRIDE is flexible in addressing further applications just by defining an appropriate tagging of the target sequence since the primer design software only relies on the sequence tags that define coding regions/exons. The design of primers for the amplification of overlapping 1000 bp fragments of a genome sequence for instance, might be implemented by the generation of overlapping CODG tags of the appropriate size. In this example CODG tags do not reflect the original feature of marking protein-coding sequences. Instead, the user takes advantage of the fact that GenomePRIDE is able to design primers close to tag boundaries without prior verification of the tags. So far, GenomePRIDE has been used successfully for whole genome-specific PCR, amplification of genomic regions, amplification of splice isoform specific PCR fragments, and in RNAi experiments supported by experimental data derived from large-scale projects. Applications in single nucleotide polymorphism detection are currently under investigation. Future developments will also include further optimization of PCR fragments for RNAi experiments.

GenomePRIDE is a fast and flexible software for the design of PCR primers and/or long oligomers that can easily be adapted to many different applications. The integration into the Staden package provides a convenient graphical user interface that greatly facilitates the management of large-scale projects. With its high record of success, GenomePRIDE reduces the time and costs needed for the implementation of whole genome approaches significantly.

**AVAILABILITY**

To academic institutions GenomePRIDE is available for a fee of 150 Euro that is intended to cover our costs of distribution and maintenance (see http://pride.molgen.mpg.de/genomepride.html). The licence includes precompiled executables (Linux, Solaris, OFS) as well as the entire source code together with scripts facilitating reccompilation.

**ACKNOWLEDGEMENTS**

We would like to thank S. Wiemann and F. Wilmer for fruitful discussion and their patience in testing the software. Many thanks to B. Beckmann and B. Koch who supported the laboratory work of the *Drosophila* project. This work was supported by a grant from the German Human Genome Project (DHGP) and partly done at the German Cancer Research Center (DKFZ).

**REFERENCES**


