High-throughput Primer and Probe Design for Genome-Wide DNA Methylation Study Using PRIMEGENS

Gyan Prakash Srivastava
Computer Science Department and Christopher S. Bond Life Sciences Center
University of Missouri, USA

Garima Kushwaha
Informatics Institute and Christopher S. Bond Life Sciences Center
University of Missouri, USA

Huidong Shi
MCG Cancer Center
Medical College of Georgia, USA

Dong Xu
Computer Science Department, Informatics Institute, and Christopher S. Bond Life Sciences Center
University of Missouri, USA

1 Introduction

Research in the field of DNA sequencing has generated tremendous amounts of genomic sequences. These DNA sequence information have often been used to enhance understanding of genomic changes and genetic diseases. In addition to studying DNA sequences and genetic changes, the study of epigenetic changes that do not involve DNA sequence has also become one of the major focus areas in medical research. DNA methylation, primarily methylation of cytosine at CpG dinucleotides, is the most widely studied epigenetic modification and is known to have profound effects on gene expression. It is involved in embryogenesis, genomic imprinting [Razin & Cedar, 1994], X-chromosome inactivation [Jaenisch & Bird, 2003], and development of diseases [Li, 2002], particularly various types of cancers [Jones & Takai, 2001]. Methylation studies in cancer research have identified methylation of hundreds of CpG islands (methylation susceptible sites) in a tumor cell. Early diagnostic techniques and therapeutics based on epigenetic strategies have also been considered for cancer treatment and diagnosis [Woonbok et al., 2008]. DNA methylation also plays a critical role in the development of many other human diseases, such as neuro-developmental disorders, cardiovascular diseases, type-2 diabetes, obesity and infertility [Van et al, 2007]. DNA methylation is also involved in normal cellular processes, including gene regulation, DNA-protein interaction, and cellular differentiation [Shiota, 2004, Eden & Cedar, 1994].

DNA methylation mainly refers to the addition of the methyl group to the carbon-5 position of cytosine (5MeC) at CpG dinucleotides present in DNA sequences. In human somatic cells, 5MeC accounts for 1%
of total DNA bases and affects 70 - 80% of all CpG dinucleotides in the genome [Ehrlich et al., 1982]. Genomic regions of 300-3,000 base pairs that have very high CpG dinucleotide density are known as CpG islands. There are around 29,000 CpG islands identified in the human genome [Venter et al., 2001, Esteller & Herman, 2002]. Approximately 70% of gene promoters are associated with these CpG islands. CpG islands in promoter regions are mostly unmethylated in a normal cell at all stages of development and in all tissue types; these are often associated with active gene transcription [Antequera & Bird, 1993]. However, the methylation profile of a cell can be controlled during its development. The origin of genomic DNA methylation and factors dictating methylation status of genes remain a mystery [Feng et al., 2007]. Most DNA methylations take place at CG-dinucleotides of CpG islands but some studies have also found non-CG methylation on DNA sequences associated with highly expressed genes [Clark, 2007, Flanagan & Wild, 2007].

DNA methylation patterns in cancer cells show major variations compared with normal cells and are usually associated with transcriptional inactivation of tumor suppressor genes [Das & Singal, 2004, Fujikane et al., 2009, Kistensen et al., 2009]. Aberrant methylation patterns in cancer cells typically consist of hypomethylation involving repeated DNA sequences, such as long interspersed nuclear elements and hypermethylation at CpG islands [Ehrlich, 2002, Libault et al., 2009]. There are numerous tumor suppressor genes that are known to be hypermethylated; these include genes involved in cell cycle regulation (p16INK4a, p15INK4a, Rb, and p14ARF), DNA repair (BRCA1 and MGMT), apoptosis (DAPK and TMS1), drug resistance, detoxification, differentiation, angiogenesis, and metastasis. Hypomethylation is common in solid tumors, such as metastatic hepatocellular [Lin et al., 2001, Libault et al., 2009] and cervical cancer [Kim et al., 1994, Libault et al., 2009], prostate tumors [Bedford and Helden, 1987, Libault et al., 2009], and in hematologic malignancies such as B-cell chronic lymphocytic leukemia [Ehrlich, 2002, Libault et al., 2009]. Hypomethylation participates in oncogenesis by activating oncogenes such as c-MYC and H-RAS and activation of latent retrotransposons [Bedford and Helden, 1987, Feinberg and Vogelstein, 1983], thereby causing chromosome instability [Tuck et al., 2000]. Hypomethylation of long interspersed nuclear element DNA causes transcriptional activation and is found in many types of cancers, such as urinary bladder cancer [Jurgens et al., 1996]. Both genome-wide and candidate gene approach for methylation studies predict that each tumor or cancer type has its own set of cell-type specific genes associated with methylation. Different panel of genes can be used as genetic markers to detect aberrant methylation in certain cancers [Shivapurkar & Gazdar, 2010], such as lung and breast cancers [Brook et al., 2009], to diagnose their early stage of development [Woombok et al., 2008]. DNA methylation, together with covalent modification of histones, is known to alter chromatin packing known as “chromatin remodeling”, which changes the accessibility of DNA to cellular machinery or proteins that help transcribe the underlying DNA sequence [Flanagan and Wild, 2007].

Since the disruption of DNA methylation has become a major hallmark of human cancer, precise mapping of methylation patterns in CpG islands [Shen et al., 2007] has become an important tool in understanding cancers and other diseases. For the past several years, an international consensus has emerged in the epigenetics research community for the need of an organized Human Epigenome Project (HEP) aimed at generating a high-resolution DNA methylation map of the human genome in all major tissues [Eckhardt et al., 2006, Rakyan et al., 2004]. The HEP intends to provide a “reference epigenome” by resequencing different normal tissues and adding 5-methylcytosine to the DNA sequencing datasets (http://nihroadmap.nih.gov/epigenomics). The pilot HEP in Europe utilized direct sequencing of bisulfite polymerase chain reaction (PCR) products to provide single methyl-cytosine resolution mapping of thousands of amplicons [Eckhardt et al., 2006]. Moreover, an innovative massively parallel sequencing-by-synthesis method (454-sequencing) has been applied in ultra-deep bisulfite sequencing analysis of multiple
tumor methylome [Rakyan et al., 2004, Taylor et al., 2007].

The successful implementation of any highly parallel sequencing approach depends on the use of automatic primer design program [Liard, 2003] capable of performing genome-wide scans for optimal primers for high-throughput genomic sequences [Zhang and Smith, 2010], such as in silico bisulfite-treated human genome sequences. Several methods have been proposed to address this issue partially. There are various tools available for probe and primer design [Hillier & Green, 1991, Hydman et al. 1996, Haas et al., 1998, Herman et al., 1996, Proutski & Holmes, 1996, Li et al., 1997, Pattyn et al. 2006, Yamada et al., 2006, Tsai et al., 2007] with a few web servers [Aranyi & Tusnady, 2007], but most of them lack stringent checks for cross-hybridization of primers and probes. MethPrimer [Marshall, 2004] and PerlPrimer [Li et al., 2002] transform target sequences according to the bisulfite treatment for primer design. However, these methods do not provide a mechanism to detect non-specific amplification in bisulfite PCR. Bisearch [Tusnady et al., 2005, Aranyi & Tusnady, 2007] provides an important feature of similarity search for potential non-specific PCR products with selected primer pairs on a bisulfite-treated genome. It uses a simple string-matching search method to detect potential cross-hybridization of a designed primer pair. The string-matching search can then find an exact match for the primer; however, it cannot detect highly similar sequences (e.g., with 1 nt mismatch) in the genome to the primer, which could also be potential binding site for the primer. In addition, this method is not practical or suitable for analyzing primer pairs at mis-priming sites in high-throughput primer design [You et al., 2008, Arvidsson et al., 2006], which is required for highly parallel sequencing system to develop high-throughput, large-scale bisulfite genomic sequencing. To address this issue, an efficient method has been integrated into the second version of PRIMEGENS software [Xu et al., 2002], i.e., PRIMEGENS-v2 [Srivastava et al., 2008]. PRIMEGENS is built on third-party, open-source software tools, such as Primer3 [Rozen & Skaletsky, 2000] and BLAST, and includes various new features for genome-scale primer design. It is widely used and cited in the research community [Liu et al., 2003, Hilson et al., 2004, David, 2005, Bertone et al, 2006, Leparc et al., 2009, Lemoine et al., 2009, Libualt et al., 2009], and is available as a standalone tool to run under both Linux and Windows or as a web-based tool named PRIMEGENS-w3 (http://primegens.org/w3/index.cgi).

2 PRIMEGENS-v2 Algorithms

PRIMEGENS-v2 can be used for large-scale primer and probe design for various applications, such as PCR, DNA synthesis, qRT-PCR (gene expression), and targeted next-generation sequencing (454, Solexa, Agilent Sure-select technology, etc.) for normal and bisulfite-treated sequences. A unique feature of PRIMEGENS-v2 is that it allows users to select various types of input formats and primer design algorithms to design primers and probes customized for their experimental requirements. There is currently no other software tool that allows flexibility to incorporate various types of design constraints and customized algorithmic flowcharts suitable for different applications for PCR, microarray, and targeted sequencing.

PRIMEGENS-v2 designs primers using three different algorithms: 1) Sequence-Specific Primer Design (SSPD), 2) Fragment-Specific Primer Design (FSPD), and 3) Probe-Specific Primer Design (PSPD). It also has probe design capacity, simply called Probes Design algorithm. Detailed descriptions of each of these algorithms are provided in the following subsections.

2.1 Sequence-Specific Primer Design

Sequence-Specific Primer Design (SSPD), as shown Figure 1, is the most basic algorithm of PRIMEGENS-v2. This algorithm is used to design primers for each sequence given by the user in the query input file.
against any alternate potential hybridization with any of the sequences given in the database input file. In its first step, SSPD designs different primers for the query sequence using a third-party program, Primer3. It uses the standalone executable of Primer3 to design hundreds of primer pairs scattered all over the input query sequence by Primer3 and then retains only unique ones from this set. It further selects a handful of specific primers from the resulting multiple sets of primers. Based on the fact that stringency is more critical at the 3’-end than at the 5’-end of the primer [Onodera & Melcher, 2004], it first selects 15mer oligonucleotides from the 3’-end of each primer designed by Primer3 and then retains only the unique ones from this set. This step helps optimize the next step and allows PRIMEGENS-v2 to run faster by reducing the BLAST runs against database. The second step of SSPD is to perform gapless alignments of these 15mers against database file to search for all the potential non-specific PCR-products. For this, it uses Megablast (http://www.ncbi.nlm.nih.gov/blast/megablast.shtml) for alignment and is optimized for detection of genomic regions that are highly similar to query oligo (15mer) in the database sequences provided. Finally, SSPD selects primers that have no or minimum cross hybridizations.

As shown in Figure 2, Fragment Specific Primer Design (FSPD) is another primer-design algorithm used by PRIMEGENS-v2. This algorithm is useful in covering long sequences, such as CpG islands or other regions of interest, thereby allowing PCR to cover the whole region using multiple primers distributed uniformly across the entire region. It first breaks down the whole query sequence into small overlapping fragments and then designs primers for each of these fragments. For this, a user should provide the length of each fragment into which the query sequence will be divided (fragment length) and the overlap between two fragments to design associated primers (fragment overlap). For each fragment sequence, it functions in a manner similar to the SSPD algorithm in terms of designing a primer pair, before finally providing multiple primer pairs for a long query sequence.

### 2.2 Probe-Specific Primer Design

Probe-Specific Primer Design (PSPD), as shown in Figure 3, is another PRIMEGENS-v2 algorithm used to design primers. It selects the gene-specific fragments (probes) to design primer pairs. The two major
computational tasks in this algorithm work in a reverse order from that of SSPD. It first identifies a gene-
specific fragment (probe) and then designs primers for the target fragment (or probe). To determine gene-
specific fragments, it carries out a heuristic BLAST search for each of the query sequence against all the
sequences in the database file to identify homolog sequences. It then performs multiple sequence alignment
(star alignment) between the query and each of its homolog using Dynamic Programming. Based on the
alignment, PSPD tries to find the sequence fragments that do not align with any other sequence and selects
them as gene-specific fragments. Sequences that are unique themselves (no BLAST hit) are directly used as
gene-specific fragments. In its second task, PRIMEGENS-v2 uses Primer3 to design PCR primers for gene-
specific fragments (or probes) of the query sequence found in the first step of PSPD, instead of the whole
query sequence provided. Primer3 gives primers for PCR amplifications based on user-specified parameters
(e.g. primer size, melting temperature, GC content, self-complementarity, and so on).

Once the primer-pair design using Primer3 is accomplished and BLAST hits for each 3-end unique oligos are recorded, the model checks for potential amplicons amplified by designed primers. Only a certain order and orientation of primer binding can successfully amplify a sequence, and even if a primer pair binds in correct order and orientation, the amplicon size should fall within a desired range. Therefore, PRIMEGENS-v2 checks each of these primer selection constraints for the potential cross-hybridization of the designed primers and probes. For example, a primer pair showing potential cross-hybridization with an amplification size of more than 10,000 bases other than target region can be safely ignored for PCR experiment. Figure 4 shows an example of a primer oligo BLAST hit, which results in successful or failed amplification in PCR. Out of four different cases of primer pair hits, only a single case results in successful amplification. Furthermore, it occurs only if the amplicon size is less than the maximum possible amplicon size (the default: 10,000 bases in PRIMEGENS-v2) for PCR.

Finally, after the step of hybridization restriction, a set of generated primers is further reduced to allow
for only those primer pairs that have potentially good amplicon based on primer quality and user-specified
product size range. Out of the final designed primer pairs, top primer pairs having least number of cross-
hybridization are reported as the final output.

Figure 2: Flowchart for the FSPD algorithm.
2.3 Probe Design Algorithm for Target Enrichment

The probe design algorithm is used to search for sequence-specific probes with no BLAST hits against database sequences. This probe design has been used for targeted sequencing, including Agilent sure-select technology with next-generation sequencing. Target enrichment or targeted resequencing is useful when researchers are interested in sequencing only a certain portion of genomic region; for example only translated regions of the human genome. This technique is also known as “DNA capture” or “genome partitioning”. One of the most common platforms for target resequencing is the Agilent SureSelect platform, which allows users to capture a subset of a genome and wash away the rest of genome that can be input to any resequencing technology. The Agilent target capturing technology replaces other labor-intensive methods, such as the PCR technique, which are major bottlenecks for next-generation sequencing output. To capture certain regions of a genome rather than a whole genome, PRIMEGENS-v2 only allows users to design probes to capture those regions. This algorithm takes users inputs of the genomic region of interest in the FASTA format and designs probes (either single or multiple uniformly distributed probes for each region). The algorithm uses BLAST to ensure that it outputs only those probes that are specific to the target region and hybridize nowhere else.

To design gene-specific probes, PRIMEGENS-v2 uses a sliding window protocol to find fragments of user-given length by the PRIMEGENS-v2 variable PROB_LENGTH the window size. This sliding window moves over the length of PROBE_PERIOD (“Probe shift” in Figure 5). It looks for a probe in every genomic region given by the PRIMEGEN-v2 parameter called the PROBE_REGION (“Inter-probe distance” in Figure 5). For probe design, users can also provide their preferences in terms of maximum and minimum GC content in the probe using other PRIMEGENS-v2 variables, including MAX_GC, Min_GC, and maximum allowable BLAST hits for probes.

3 PRIMEGENS-v2 Usage

PRIMEGENS-v2 has a simple operation protocol consisting of two basic steps: (1) preparation of data files, such as list of target sequences for primer/probe design and relevant database for cross-hybridization
check in the FASTA format; and (2) selection of primer and probe design algorithm, as well as the relevant parameters along with primer specifications.

3.1 PRIMEGENS-v2: Input and Output

In order to use PRIMEGENS-v2 for primer and probe design, the user has to prepare a list of target sequences in one of the formats supported by the software and provide a database to be used for cross-hybridization check. In addition, the software provides a sample configuration file containing various types of PRIMEGENS-v2 variables as well as standard variables from Primer3 and BLAST or MegaBLAST. The configuration file also provides a description for each variable, allowing the user to select appropriate values that are customized for his or her application.

3.1.1 Input: Target Sequences

In order to use PRIMEGENS-v2 for primer and probe design, a user must provide a list of target sequences in the FASTA format, that is, each gene name should start with a “>” sign. The software also allows users to only provide gene names without their actual nucleotide sequence as long as these sequences can be retrieved from the database. The user can then provide the chromosome position of genes from any genome for primer design, for example > chr21:33031597-33041570. In this case, a user has to provide the whole genome as the database. The genome must be in the chromosomal files, and there must be one file for each sequence containing the full sequence of the corresponding genome.

PRIMEGENS-v2 provides users with four formats to input their target genes without explicit sequences, including gene name/identifier along with description, gene name/identifier, chromosome position with description, and chromosome position (see Table 1(A)). The name/identifier could be used as gene name/identifier as long as the same name/identifier is used in the database, which must be in the FASTA format.
Figure 5: Flowchart for the probe design algorithm.

PRIME-GENS-v2 uses this database and searches for the gene name/identifier in the database to retrieve actual sequence for each query input provided by the user.

Table 1(B) shows the same format as Table 1(A), but with actual sequences for each query input. In this case, the user is not required to follow any constraint on gene name/identifier in terms of matching one in the database. If users want to separate annotations with gene name/identifier, they can do so by selecting the format showing the description. When the chromosome position is given, it is still considered as gene name/identifier unlike the previous case without sequence, where the actual sequence is retrieved using this location information.

3.1.2 Input: Database (Sequence Pool/Genome)

In the second input file(s), the user has to provide PRIMEGENS-v2 a database consisting of all the sequences against which cross-hybridization has to be checked for each primer/probe designed. Based on number of hybridization, it keeps only those primers with no or minimum hybridization to the input query sequence.

3.1.3 Input: Parameters (Configuration File in the Stand-alone Version)

In addition to input query sequence and database for cross-hybridization check, the user is required to enter values for various PRIMEGENS-v2 parameters, such as primer design algorithm, query sequence format, database format, and various primer-design related parameters. These parameters are provided through the configuration file “config.txt” in PRIMEGENS-v2. To help users further, the software includes a sample configuration file in its “include” directory. The file name for this configuration file is fixed and must be named “config.txt”. The user can copy the sample configuration file and modify various parameters customized for their experimental conditions. Once a command is executed, PRIMEGENS-v2 searches for the configuration file in the user-provided output directory; if there is no configuration file in the output directory, PRIMEGENS-v2 will search for the same in the current directory. If it fails to retrieve the configuration file, PRIMEGENS-v2 will use the default sample configuration file provided by the software in the “include”
G. P. Srivastava, G. Kushwaha, H. Shi & D. Xu

**Input target sequence ID format without DNA sequence**

<table>
<thead>
<tr>
<th>&gt; S11529022 description ...</th>
<th>&gt; S11417986</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; S22334604 description ...</td>
<td>&gt; S11417988</td>
</tr>
<tr>
<td>&gt; S11417968 description ...</td>
<td>&gt; S11528870</td>
</tr>
<tr>
<td>&gt; S11431057 description ...</td>
<td>&gt; S11417995</td>
</tr>
<tr>
<td>&gt; S11520763 description ...</td>
<td>&gt; S11418609</td>
</tr>
</tbody>
</table>

**Input target sequence location without DNA sequence**

<table>
<thead>
<tr>
<th>&gt; chr21:170390-171698 description ...</th>
<th>&gt; chr21:170390-171698</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; chrX:222396-222882 description ...</td>
<td>&gt; chrX:222396-222882</td>
</tr>
<tr>
<td>&gt; chr14:36539-40265 description ...</td>
<td>&gt; chr14:36539-40265</td>
</tr>
<tr>
<td>&gt; chr16:36539-38579 description ...</td>
<td>&gt; chr16:36539-38579</td>
</tr>
</tbody>
</table>

**A. Without actual sequences**

**Input target sequence ID format with DNA sequence**

<table>
<thead>
<tr>
<th>&gt; S11529022 description ...</th>
<th>GACGCGCAGACGTGATAAAGCTG ...</th>
<th>&gt; S11417986</th>
<th>GACGCGCAGACGTGATAAAGCTG ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; S22334604 description ...</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
<td>&gt; S11417988</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
</tr>
<tr>
<td>&gt; S11417968 description ...</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
<td>&gt; S11528870</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
</tr>
<tr>
<td>&gt; S11431057 description ...</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
<td>&gt; S11417995</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
</tr>
<tr>
<td>&gt; S11520763 description ...</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
<td>&gt; S11418609</td>
<td>GACGCGCAGACGTGATAAAGCTG ...</td>
</tr>
</tbody>
</table>

**Input target sequence location with DNA sequence**

<table>
<thead>
<tr>
<th>&gt; chr21:170390-171698 description ...</th>
<th>GACGCGCAGACGT ...</th>
<th>&gt; chr21:170390-171698</th>
<th>GACGCGCAGACGT ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; chrX:222396-222882 description ...</td>
<td>GACGCGCAGACGT ...</td>
<td>&gt; chrX:222396-222882</td>
<td>GACGCGCAGACGT ...</td>
</tr>
<tr>
<td>&gt; chr14:36539-40265 description ...</td>
<td>GACGCGCAGACGT ...</td>
<td>&gt; chr14:36539-40265</td>
<td>GACGCGCAGACGT ...</td>
</tr>
<tr>
<td>&gt; chr16:36539-38579 description ...</td>
<td>GACGCGCAGACGT ...</td>
<td>&gt; chr16:36539-38579</td>
<td>GACGCGCAGACGT ...</td>
</tr>
</tbody>
</table>

**B. With actual sequences**

**Table 1:** List of all four formats without or with actual sequences supported by PRIMEGENS-v2.

directory. The detailed description of each parameter is provided in the configuration file in lines starting with the “#” character as comments. Most of the parameters have been set to a default value as standard parameters for the best primer design. These parameters are divided into different sections described below.

1. **Parameters for the primer-design algorithm type** This section of parameters allows the user to choose primer and probe design algorithms out of the three primer and one probe design algorithms supported by PRIMEGENS-v2. The default algorithm for primer design is SSPD.

2. **Parameters required for the BLAST and Primer3 programs** In this section, the user can set parameters for MegaBLAST to search for cross-hybridization of primers in the database sequences and the desired characteristics of the primer used by Primer3 to design primers, including melting temperature and primer length, among others.

3. **Parameters required for the FSPD program** This section specifies parameters required for the FSPD algorithm if chosen. This comprises the length of each fragment into which the query sequence is divided (fragment length) and the overlap between two fragments to design associated primers (fragment overlap).

4. **Parameters required for the PSPD program** This section specifies parameters required for the PSPD algorithm if chosen. These parameters include gene-fragment length and the maximum allowed se-
160 Chapter 11 – High-throughput Primer and Probe Design for Genome-Wide DNA Methylation Study

Table 2: Sample Excel output file generated by PRIMEGENS-v2 for designed alternate primer pairs (with relatively high number of cross-hybridizations).

<table>
<thead>
<tr>
<th>&gt;chr1:20366370-20376857</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) CAGAAAAGCTGTACACGCCA [190] TGTACGCATGTGTGTGTGGT [359] psize 170 hbrdn 163</td>
</tr>
<tr>
<td>2) CAGAAAGGAGCGAGCAGTCT [166] TGTACGCATGTGTGTGTGGT [359] psize 194 hbrdn 171</td>
</tr>
<tr>
<td>3) AACGTCTGTGTGACTCAGTGC [117] TGTACGCATGTGTGTGTGGT [359] psize 243 hbrdn 176</td>
</tr>
<tr>
<td>4) AACGTCTGTGTGACTCAGTGC [117] TGTGTGTGGTGTGTGTGCAT [349] psize 233 hbrdn 189</td>
</tr>
<tr>
<td>5) AACGTCTGTGTGACTCAGTGC [117] GTACGCATGTGTGTGTGGTG [358] psize 242 hbrdn 250</td>
</tr>
<tr>
<td>6) CAGAAAGCTGTACACGCCA [190] GTACGCATGTGTGTGTGGT [358] psize 193 hbrdn 251</td>
</tr>
<tr>
<td>7) CAGAAAGGAGCGAGCAGTCT [166] GTACGCATGTGTGTGTGGT [358] psize 193 hbrdn 265</td>
</tr>
<tr>
<td>8) CAGAAAGGAGCGAGCAGTCT [166] GTGTACGCATGTGTGTGTGG [360] psize 171 hbrdn 310</td>
</tr>
<tr>
<td>9) CAGAAAGGAGCGAGCAGTCT [166] GTGTACGCATGTGTGTGTGG [360] psize 195 hbrdn 329</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>&gt;chr1:235833063-235833326</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) TTAGACTCTAGTCTGGCACGG [131] TAGCACCATTCCACAAGGC [351] psize 221 hbrdn 10</td>
</tr>
<tr>
<td>2) CTTGACTCTAGTCTGGCACGG [130] TAGCACCATTCCACAAGGC [351] psize 222 hbrdn 10</td>
</tr>
<tr>
<td>3) TCTAGTCTGGCACCGTGAAAG [136] TAGCACCATTCCACAAGGC [351] psize 216 hbrdn 12</td>
</tr>
<tr>
<td>4) CTTAGTCTGGCACCGTGAAAG [135] TAGCACCATTCCACAAGGC [351] psize 217 hbrdn 12</td>
</tr>
<tr>
<td>5) ATGACGGAGGATTTCTGCTAC [202] TAGCACCATTCCACAAGGC [351] psize 150 hbrdn 14</td>
</tr>
</tbody>
</table>

5. Parameters for Probe Design This section specifies parameters required for probe design if chosen. These parameters include probe length and allowed number of probe BLAST hits, among others.

3.1.4 Output: Best Primer Pairs

The most important output file is the Excel sheet generated by PRIMEGENS-v2, which shows the best primer pairs for each of the input query sequence provided by the user. This file also contains various types of primer pair-related information and product size from each primer pair. Table 2 shows the format of this Excel sheet output file.

3.1.5 Output: Alternate Primer Pairs

In addition to the best primer pairs, PRIMEGENS-v2 produces another file containing alternate primer pairs for each input query sequence. In case a user wants to select alternate primer pairs instead of the best primers, or if the best primer pairs fail to amplify or match other desired criteria, this file provides multiple choices for additional primer pairs for each query sequence. These primer pairs are ranked based on number of potential cross-hybridizations. Table 3 shows the format of this output file.

3.1.6 Output: Failed Sequences

PRIMEGENS-v2 also generates one file containing all FASTA formatted sequences, in which the primer design failed. There could be various reasons for the failure, including design failure from the Primer3 program. In this case, the user can analyze these sequences further and re-run PRIMEGENS-v2 only for failed sequences using modified parameters.
3.1.7 Output: Probe for Target Sequencing

In addition to three standard output files, PSPD generates an additional output file, i.e., gene-specific fragments (recorded with name of the query file followed by “uniseg.txt”). This file contains gene-specific fragments (probe) for each input query sequence that PSPD finds. These are the gene-specific fragments that PSPD ultimately uses to design primers for their corresponding query sequences.

4 PRIMEGENS-w3: Web Server for PRIMEGENS-v2

PRIMEGENS-w3 is the Web Server version of PRIMEGENS-v2. It is meant to automate high-throughput primer and probe design. PRIMEGENS-w3 allows users to utilize our computing resources in conducting PCR primer and probe designs for various applications. Having this available online removes the necessity for users to install the software, and they can instead use a web server and with stored database on the server in performing cross-hybridization checks.

PRIMEGENS-w3 consists of two basic steps: 1) uploading data files (PCR template files for primer design and optional database for cross-hybridization check); and 2) identifying primer-design specifications for job submission and execution. PRIMEGENS-w3 allows user to select any of the algorithms for primer design similar to its stand-alone version. The task of designing primers and probes using the web server also requires the same two types of inputs. One is the query file having the sequence for which primers/probes need to be designed, and the other is the database file having all the sequences that are present in the PCR reaction. Sequences in the database files are used for PRIMEGENS-w3 to check for any potential cross-hybridization and thereby select primer/probes that are specific to the query sequence. The user can upload their own customized database file or use available genomes supported by the server. We do not recommend user to upload large database like any genome. Any new genome can be made available on the web server based on user request. PRIMEGENS-w3 also provides different sample data for both query and database sequences to enable users to test primer/probe design.

The next stage of PRIMEGENS-w3 is to provide all input parameters for primer design. Input parameters on this page of the server are divided into same five sections as described in section 3.1.3. All parameters have been set to the default values for best primer design. The default values are visible on the browser and they can be changed. After running PRIMEGENS-w3, the server will show the link to all result files generated by the tool. PRIMEGENS-w3 also generates files similar to those by the standalone version as explained in Sections 3.1.4 - 3.1.7. These files can be viewed within the browser or downloaded on the users local computer. In case the PRIMEGENS-w3 execution takes long time to complete due to large size of input files, the server provides user with a link to be used later to check for the updated result files. Along with this, a user can also provide an email address (optional) to enable PRIMEGENS-w3 to send them an email notification for job completion and the link for the result files. Figure 6 shows the first page of the PRIMEGENS-w3 server. As can be seen, this page shows various options for input files required by PRIMEGENS-w3. Figure 7 shows the second page of PRIMEGENS-w3 containing the forms wherein the user can indicate the desired parameters.

5 Application of PRIMEGENS-v2 for DNA Methylation Study

PRIMEGENS-v2 has several useful features that can be utilized in designing high-quality primers and probes specifically for use in DNA methylation study in association with any disease or other biological phenotype.
5.1 Primer and Probe Design for Bisulfite-Converted Sequences

There are various techniques that can precisely map cytosine-methylation sites in a DNA sequence. Of these techniques, the bisulfite conversion-based method is the most widely used. In this method, DNA is first denatured to a single-stranded DNA and then treated with sodium bisulfite to convert cytosine to uracil. Through this reaction, all unmethylated cytosines are converted to uracil by deamination, leaving the methylated cytosines unconverted. After this modification step, DNA is amplified using two different PCR methods: methylation-specific PCR (MSP) and bisulfite-sequencing PCR (BSP). In MSP, the methylation status of the DNA is assessed by designing two sets of primer pairs with at least one CpG site in each pair. One pair is specifically designed to amplify methylated DNA (M pair), while the other pair amplifies the unmethylated DNA (U pair). Therefore, for each DNA sequence, two PCRs are carried out using each primer pair. If a sequence becomes amplified by primers with methylated DNA (M pair), it indicates the presence of methylation of CpG site within the primer sequence. On the other hand, if it becomes amplified by the other pair for unmethylated DNA (U pair), this represents that no methylation of CpG site occurred inside the primer sequence. Bisulfite-specific PCR (BSP) is another method used to study DNA methylation after the bisulfite treatment of the DNA sequence. Primers used for BSP contain no CpG sites; thus, both methylated and unmethylated DNA sequences can be amplified. In a PCR product, identifying cytosines indicates methylation, while identifying thymine at the same position indicates the lack of methylation within the amplified region.

PRIMEGENS-v2 can design PCR primers for bisulfite conversion-based PCR reaction. It requires the input of bisulfite-converted query sequences and four variants of the database against which query sequence are checked for cross-hybridization. To generate four genome variants after bisulfite conversion, all the cytosine (C) sites in the original sequence are converted into thymine (T), except for places where cytosine precedes guanine (G), known as methylation of CG. In other words, for each chromosome sequence, four chromosomal files are generated as the model of the bisulfite-treated sequences: (1) bisulfite-methylated
forward sequence, (2) bisulfite-methylated reverse sequence, (3) bisulfite-unmethylated forward sequence, and (4) bisulfite-unmethylated reverse sequence. As an example, suppose a human genome fragment has a nucleotide sequence "agctagccagtcga", this fragment is then modified to generate four variants as follows:

- agctagccagtcga - original DNA sequence
- agttagttagttga - unmethylated forward DNA sequence
- agttagttagctga - methylated forward DNA sequence
- ttgattggttagtt - unmethylated reverse complementary DNA sequence
- tcgattggttagtt - methylated reverse complementary DNA sequence

The PRIMEGENS-v2 configuration file provides the following tags dedicated for bisulfite conversion-based PCR and MSP:

- **BISULFITE_MODE = 1 / 0** (1 for bisulfite-conversion based PCR and 0 for normal PCR),
- **PRIMER_CG_MAX = -1 / +N** (-1 = ignore this rule, +N = allows max “N” CGs within primer oligo).

### 5.2 Primer and Probe Design around the Transcription Start Site

“Primer design covering TSS” is a feature of PRIMEGENS-v2, which helps the designed primers to cover the region around a transcription start site (TSS) of any gene.

#### 5.2.1 Uniform Coverage for a Sequence Region

For probe and primer design around the TSS of a gene, the FSPD algorithm is used to divide the query sequence into different fragments, including the surroundings of the TSS position for the gene present in
5.2.2 Coverage of Specific Region around the TSS

For covering specific region around the TSS of a human gene, the user is only required to provide gene symbols for which the primer design is required. PRIMEGENS-v2 is capable of extracting their respective TSSs from the UCSC Genome Database (currently for March 2006 assembly). Users can also provide directions from the TSS or region of interest for primer design. Based on this information, PRIMEGENS-v2 will extract the final query sequence to be used for the ultimate primer design. Query sequence is chosen in such a manner that it covers the TSS and the region in either side. Regions around the TSS can be defined out of three options for the primer design, including “First Exon Coverage”, “Across TSS”, and “Promoter Coverage”. The length of the query sequence to be used for the primer design is determined by another PRIMEGENS-v2 variable, i.e., QLENGTH. Another important PRIMEGENS-v2 variable is MINCOVER, which determines the extension of the query sequence in the direction opposite to the side of the coverage preference. It tries to cover either the whole or partial exon/promoter region within the region of QLENGTH. If the sum of the first exon length and MINCOVER is smaller than the QLENGTH, then the QLENGTH is changed to cover just the exon-1; thus avoiding extending the query sequence for the primer design into next intronic sequence. On the other hand, if the QLENGTH is smaller than the total of the MINCOVER region and the first exonic region, i.e., QLENGTH is unable to cover the whole exon-1, and then PRIMEGENS-v2 considers the partial coverage of exon-1. Figure 9 illustrates this feature of PRIMEGENS-v2 to design primers around the TSS of candidate genes based on users selections.

5.3 Primer and Probe Design in or around CpG Island near the Transcription Start Site

“Primer design covering CpG Island” is one of the unique features of PRIMEGENS-v2, which can be used to study methylation patterns of various oncogenes and tumor suppressor genes. This feature designs primers
for genes that have CpG islands present in close proximity to their respective TSSs. Primers can be designed to amplify genes whose expressions are suspected to be influenced by nearby CpG islands. In this approach, locating CpG islands is determined by a PRIMEGENS-v2 variable, VICINITY. PRIMEGENS-v2 looks for the presence of CpG islands in the region of VICINITY, in both sides of a genes TSS. It tries to cover either the whole or partial CpG islands in the region of the VICINITY value. CpG islands can lie to the left, right, or across the TSS. In addition, users can also provide their direction or region of interest to enable PRIMEGENS-v2 to look for CpG islands. By default, CpG islands present in any of the three possible positions are considered, and the distance between a CpG island and TSS, if found at one or the other side of TSS are also reported. Based on this information, PRIMEGENS-v2 extracts the final query sequence for the primer design. The length of the query sequence to be used for primer design is also determined by PRIMEGENS-v2 variables QLENGTH and MINCOVER (see Section 5.2.2). The FSPD algorithm is recommended for this type of primer design due to the long length of CpG islands. Figure 10 shows this feature of PRIMEGENS-v2, illustrates a design primer for genes surrounded by CpG islands. Thus far, this feature has been used and supported for human genome and their TSS and CpG locations. Other supported genomes will be used and tested in the future.

5.4 Probe design for Targeted Sequencing to Screen DNA Methylation Patterns of All CpG Islands in the Human Genome

We applied PRIMEGENS-v2 in ultra-long oligonucleotides for a hybridization-based capture of CGIs and their flanking sequences from a randomly fragmented shotgun genomic DNA library (Figure 11). The long oligos (160bp) were synthesized by Agilent Technologies to capture the DNA representing 28,226 CGIs
Figure 10: Amplified CpG islands within the vicinity of a gene's TSS. (A) shows the presence of CpG island at the left side of the TSS (gradient blue) lying inside the region of VICINITY. Here, the query sequence has been picked of the length of QLENGTH value covering the MINCOVER region at the opposite side of the TSS (right side), the TSS itself, and portion of the CpG on the left. (B) shows the presence of CpG island across the TSS (gradient blue). Here, the query sequence has been picked of the length of QLENGTH value equally covering both sides of the TSS. (C) shows the presence of CpG island at the right side of the TSS (gradient blue) with its start position lying inside the region of VICINITY at the right side of the TSS. Here, the query sequence has been picked of the length of QLENGTH value covering the MINCOVER region at the opposite side of the TSS (left side), the TSS itself, and portion of CpG at the left.

(sizes ranging from 200bp to 5000bp) dispersed throughout the genome. The single-strand DNA oligos were converted to biotin-labeled RNA baits based on the protocols established by Gnirke et al. [Gnirke et al., 2009]. The RNA baits were used to capture the CpG island DNA from a shotgun-sequencing library consisting of breast cancer cell line DNA. The captured DNA was then treated with bisulfite and converted into DNA sequencing libraries using PCR in a single-tube format. The CGI-enriched, bisulfite-modified genomic DNA library was then sequenced using next-generation sequencer Solexa. As can be seen from the results, the probe design captured most of the intended target sequences (Figure 12).

5.5 Primer and Probe Design around the Maximum Cut-site Region

PRIMEGENS-v2 can also be used to search for regions with the maximum enzyme digestion sites (cut-sites) within each query sequence and design primers around these cut-sites. This ensures the presence of cut-sites in the PCR product and is very useful in MSP. PRIMEGENS-v2 searches for any of the cut-sites provided by the user in the query sequence. From each cut-site, it tries to search for the region with the maximum density of such sites in the regions that are determined by the user as another PRIMEGENS-v2 variable known as CUT_SITE_REGION. The region with the maximum cut site density is then considered as the region for TARGET_REGION variable used by Primer3 to design primers around them. Figure 13 shows how PRIMEGENS-v2 locates the maximum cut-site region in the whole query sequence and chooses the
Figure 11: Solution-based hybridization capture of CGIs and targeted shotgun bisulfite-treated sequencing. (A) Left panel illustrates the steps involved in the preparation of a complex pool of biotinylated RNA capture probes (top left) using Agilents array-based ultra-long oligonucleotides synthesis method. The whole-genome fragment input library (pond; top right) was created using Illumina methylated pair-end adaptors. (B) The hybrid-selected, enriched output library (catch; bottom) was then treated with bisulfite and amplified using Illumina PCR primers. Two sequencing targets and their respective baits are shown in red and blue. The universal adaptor sequences are gray. The excess of single-stranded non-self-complementary RNA (wavy lines) drives the hybridization. (The left panel came from Nature Publishing Group)

target region within CUT_SITE_REGION, as well as how Primer3 designs primers around the target region.

6 Conclusion

Highly specific and efficient primer/probe design has been in demand for genomic research. The presence of similar sequences in a sequence pool has often caused difficulties in the identification of unique sequences of interest using PCR or pull-down approaches. Therefore, designing primers that enable PCR to specifically amplify one sequence from a large database of sequences in a high-throughput fashion is a major challenge. PRIMEGENS-v2 offers a high-throughput approach in conducting primer and probe design for large-scale amplification and sequencing projects. It automates the design of a reliable set of primers with a minimum rate of failure due to cross-hybridization. PRIMEGENS-v2 features can also be efficiently used in studying DNA methylation. Furthermore, while other existing primer design tools usually design primers for one gene at a time, PRIMEGENS-v2 can design primers for thousands of genes (fragments) in one run.
Figure 12: A genome browser view of designed probes and captured DNA sequences. The black squares show the designed probes targeting CGIs. The varying bars indicate the sequence abundance obtained by the Solexa sequencer at the genomic coordinates.

PRIMEGENS-v2 is the only available tool that can automatically check primer specificity with customized inputs of sequences on a large scale. Therefore, PRIMEGENS-v2 provides a unique software package for the research community as compared with other available tools for primer design.

Acknowledgements

This work was supported by the National Institute of Health (Grant No. 1R01-DA025779) and the Congressionally Directed Medical Research Programs, U.S. Army Medical Research and Material Command (Grant BC062135). We would like to thank Juyuan Guo, Muneendra Ojha, Michael Wang, and Trupti Joshi for helpful discussions or technical assistance. Major computations were performed using the UMBC computing resource.

References


Scanning CUT_SITE_REGION Across Whole Sequence

**Figure 13:** The PRIMEGENS-v2 algorithm for choosing the maximum cut-site region in query sequence for the primer design. It systematically searches for segments including the cut-sites and then identifies the regions with the maximum cut-sites for primer design.


