# Universal-, Genus-specific, Species-specific Probes and Primers Design for Microbial Identification

**Junhyung Park[1] Heekyung Park[2] Eunsil Song[2] Hyunjung Jang[2] Byeongchul Kang[3] Seungwon Lee[1] Hyunjin Kim[1] Cheolmin Kim[1]**

[1]*Busan Genome Cente, College of Medicine, Pusan National University, Busan, South Korea*
[2]*Institute for Genomics Medicine, GeneIn. Co., Ltd. Busan, South Korea*
[3]*Devision of Applied Bioengineering, Dongseo University, Busan, South Korea*
*Email : jhpark98@pusan.ac.kr, dnachip@dreamwiz.com, seri1026@dreamwiz.com, biochip@dreamwiz.com, bckang@dongseo.ac.kr, user@pusan.ac.kr, asever@pusan.ac.kr, kimcm@pusan.ac.kr*

**ABSTRACT**: MIPROBE is a web-based tool for design of universal, genus-specific, and species-specific primers and probes. The main functions of MIPROBE are collection of target gene sequences, construction of consensus sequences, collection of candidate primers and probes, and evaluation of candidates by BLAST. Biologists with little computer skills can easily use MIPROBE to design large-scale universal, genus-, and species-specific primers and probes. This software is available at http://www.miprobe.com. Also detailed descriptions of how to use the program are found at this site.

## 1 INTRODUCION

Ribosomal DNA-(rDNA)-targeted primers and probes are widely used for identification of microorganisms in environmental and clinical samples [1]. rDNA genes possess highly conserved regions which are suitable as sites for PCR primers that recognize large group of organisms, as well as variable regions that provide signatures for more accurate identification. Previous investigators have usually chosen the 16S rDNA ribosomal DNA or the 16S-23S rDNA spacer region as a target for universal primers. The 16S rDNA is highly conserved, and sequences from it are now used in bacterial taxonomy. In constrast, the 16S-23S rDNA spacer region is highly variable within many species, and this variation has been used for typing clinical isolates [2]. Recently, sequence data for 23S rDNA have become available for bacterial species. This region shows more variation than 16S rDNA between important species in medicine; therefore primer and probe design using 23S rDNA might be more useful for clinical diagnosis [3].
Amplication by universal sequences in pathogenic bacterial DNA would allow rapid identification of pathogenic bacteria, and amplification of genus-specific and species-specific sequences of pathogenic bacterial DNA might be used for genotyping at the genus and species level[4].

In order to use these approaches, universal, genus-specific, and species-specific primers and probes design tools should be developed. Numerous tools for design of primers and probes are available as stand-alone programs or as web application such as PrimerMaster, PrimeArray, OligoArray, PRIDE, and Web Primer. However, most of these programs can design only a few primer sets at one time and are therefore less suitable for large-scale primers and probes design [5]. Also, none are capable of automatically performing large-scale design of primers and probes. Therefore, to design specific primers and probes, researchers must use various tools in a semi-manual manner. Moreover, outputs of sequence alignment must be checked manually, and inputs for the primer design of each sequence must be entered manually as well. This process is time-consuming [6]. Accordingly, a high-throughput approach towards automated design tools has become essential. And web-based schemes are more efficient for large-scale project. We present the methods and strategies for the web-based large-scale universal, genus-specific, and species-specific primers and probes design for pathogen identification based on 23S rDNA.

## 2 SYSTEMS AND METHODS

Results of the designed primers and probes are stored in a MySQL database. We designed a web interface and connected it to that database using Perl. This allows one to query the database in a user-friendly manner.

## 3 IMPLEMENTATION

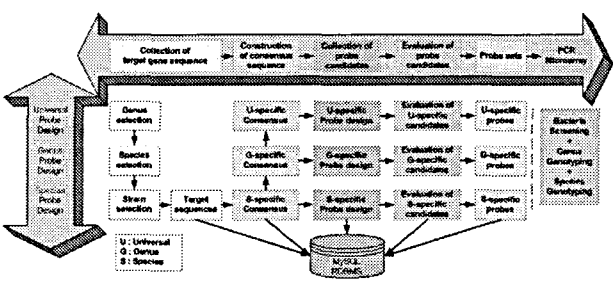

Figure 1: System architecture of MIPROBE

The system involves four major components: 1) collection of target gene sequences; 2) construction of consensus sequences; 3) collection of primer and probe candidates; 4) evaluation of candidates by BLAST.

### 3.1 Collection of target gene sequences

In large-scaled design of primers and probes, one of the most time-consuming tasks is correct collection of the designed target gene sequences. We can obtain target gene sequences from NCBI GenBank and through our own direct

DNA sequencing. An Entrez search can be used to find one or more sequences in NCBI GenBank by inputting text queries into fields that are linked to the raw sequence data.

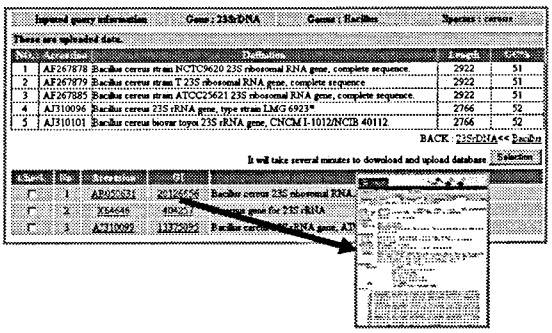

Figure 2: Collection of target gene sequences

Sometimes, we use a Boolean search(AND or OR) in order to collect more exact sequences. Nevertheless, Entrez shows target gene sequences and non-target sequences at the same time. Therefore, the results need to be checked by manually one by one. If a researcher would like to obtain hundreds of sequences against different genera and species, it is a laborious work. So we downloaded nt.gz, a compressed non-redundant nucleotide sequence file, from the FTP repository in NCBI, and have customized the nucleotide sequence list. Once the target gene, the genus, and the species are selected in order, strain information including accession, genbank id, and definition are suggested, based on the customized nucleotide sequence list. And if a researcher selects strains among the suggested list, the selected strain sequences can be downloaded from NCBI GenBank using a sequence extraction system that was developed. This component has been designed to be as user-friendly as possible, allowing a researcher to avoid laborious and time-consuming work.

## 3.2 Construction of consensus sequence

Consensus sequences are extracted from a multiple-sequence alignment using local ClustalW which is a progressive multiple-sequence alignment program. Selection of consensus sequences has been designed using a hierarchical method. Therefore, a species-consensus sequence is extracted from multiple-sequence alignment results of strain sequences, and a genus-consensus sequence is extracted from species-consensus sequences. In the same manner, a universal sequence is extracted from genus-consensus sequences.

The consensus sequence is used to collect the primers and probes candidates. If the wrong sequence is contained in running a multiple-sequence alignment, an inaccurate consensus sequence could result. So, it is important to confirm multiple-sequence alignment results before performing probe design. This component shows various types of multiple sequence alignment results and is linked to 'Phylodendron' (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html). It allows us to determine whether the consensus sequence is accurate. If the consensus sequence is determined to be inaccurate, it is possible to perform multiple-sequence alignment, excepting the wrong sequence again.

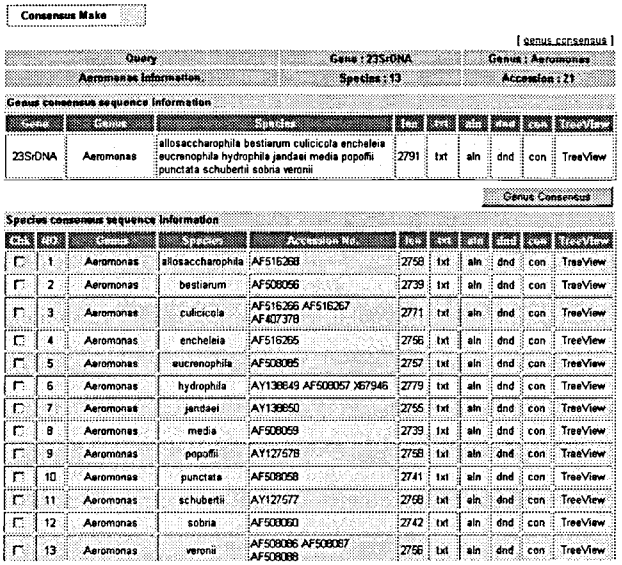


Figure 3: Construction of consensus sequence
: Example of genus consensus sequence

## 3.3 Collection of primer and probe candidates

The algorithm for collection of primer and probe candidates is divided into two distinct parts. In the first step, MIPROBE scans along the consensus sequence in a sliding window scheme in one-nucleotide shifts. It isolates conserved regions and non-conserved regions that include 'N' in the consensus sequence. In the second step, the conserved regions of the consensus sequence are used for computation of all of the relevant parameters including melting temperature, GC contents, length, thermodynamic stability, and self-complementarity.

Melting temperature is often used as an input for a primer design program, because the researcher requires a primer that will work under specified reaction conditions. One method for calculating melting temperature is the nearest neighbor method. The GC content describes the stability of the primer template duplex, because different energies are required to break apart GC pairs which have three hydrogen bonds and AT pairs which have only two.

MIPROBE accepts minimum, maximum oligo length and GC contents instead of a fixed length and GC contents. Within the specified range it can adjust the length of oligos to achieve greater specificity and uniformity among all oligos. MIPROBE is accomplished by a built-in parameters setting including Tm range, sequence complexity check and difference on Tm and etc,.

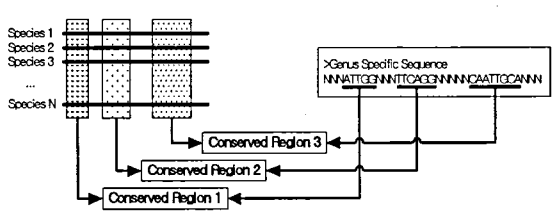


Figure 4: Choosing of conserved regions and non-conserved regions by sliding window scheme

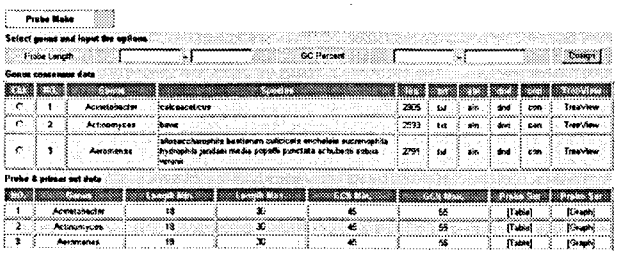

Figure 5: Collection of primer and probe candidates : Example of genus-specific design

### 3.4 Evaluation of candidates

Once the primer and probe candidates are collected, the next step is evaluation of candidate specificity between these primer and probe candidates and known sequences in the NCBI. We established up a stand-alone BLAST, which allows us to create custom-searchable database of the primer and probe candidates. We have configured the database from 133 pathogenic bacteria 23S rDNA sequences including 41 genera. Initially, a similarity search of primer and probe candidates is performed against the custom database. And each candidates is filtered according to its primer and probe type, including universal, genus-specific, and species-specific primer and probe design. Passed candidates are BLAST searched against NCBI GenBank using the automated remote BLAST method. Results of the BLAST search are parsed and updated into the database. Also, we customized the BLAST report according to the shape of the graphic taxonomy BLAST view. The graphic view report sorts the BLAST hits according to the species of the target sequence, so that all of the hits to the same organism will appear together. Within each species, the BLAST hits are sorted by score (as for the normal BLAST output). The species themselves are sorted by the strength of their strongest BLAST hit scores. In addition to the graphic taxonomy BLAST report, it shows the BLAST hits by sense strand, presence of mismatch regions, mismatch position, and presence of target gene (23S rDNA).

## 4 RESULTS AND DISCUSSION

We have designed universal, genus-, and species-specific primers and probes from 23S rDNA sequences of 41 genera including 133 pathogenic bacteria using the automated and integrated web-based tool, MIPROBE. We tested the specificity of the representative primers. The performance of the designed PCR primers was demonstrated by using human pathogenic bacteria (reference strain) including 20 genera and 100 species. Using the tested primers designed using MIPROBE, we could obtain universal, genus-specific, and species-specific amplifications. A tutorial demonstrating the collection of target gene sequences, construction of consensus sequences, collection of primer and probe candidates, and evaluation of candidates is provided on the program website.

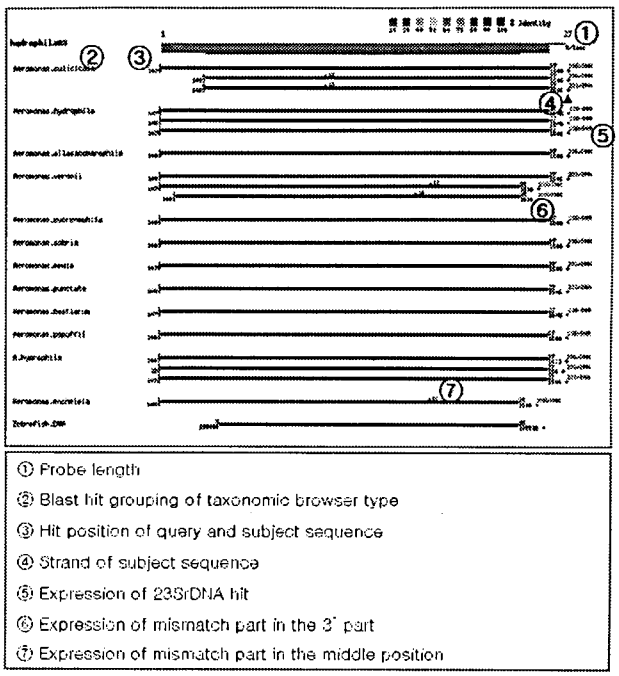


Figure 6: Candidates evaluation by graph blast

## 5 ACKNOWLEDGE

This study was supported by a grant of the International Mobile Telecommunications 2000 R&D Project, Ministry of Information & Communication, Republic of Korea.

## REFERENCES

[1] Alexander Loy, Matthias Horm and Michael Wanger. (2003) probeBase:an online resource for rRNA-targeted oligonucleotide probes, Nucleic Acids Research, 31, 514-516.

[2] H. K Park, H. J. Jang, E. S. Song, C. H. Chang, M. K. Lee, S. K. Jeong, J. H. Park, B. C. Kang, and C. M. Kim. Detection and Genotyping of Mycobacterium Species from Clinical Isolates and Specimens by Oligonucleotide Array. Journal of clinical microbiology, Vol43, no4: 1782-1788, 2005.

[3] R.M.ANTHONY. Rapid Diagnosis of Bacteremia by Universal Amplification of23S Ribosomal DNA Followed by Hybridization to an Oligonucleotide Array, Journal of clinical microbiology, 38:781-788, 2000.

[4] Kevin M.McCabe, Yao-Hua Zhang, bing-Ling Huang, Elizabeth A.Wager and Edward R.B.McCabe. Bacterial Species Identification after DNA Amplification with a Universal Primer Pair, Molecular Genetics and Metabolism, 66:205-211, 1999.

[5] Stefan Weckx, Peter De Rijk, Christine Van Broeckhoven and Jurgen Del-Favero. SNPbox:web-based high-throughput primer design from gene to genome, Nucleic Acids Research, 32:170-172, 2004

[6] Dong Xu, Guangshan Li, Liyou Wu, Jizhong Zhou and Ying Xu. PRIMEGENS:robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics, 18:1432-1437, 2002.