ABSTRACT

Motivation: Analysis of the functions of microorganisms and their dynamics in the environment is essential for understanding microbial ecology. For analysis of highly similar sequences of a functional gene family using microarrays, the previous long oligonucleotide probe design strategies have not been useful in generating probes.

Results: We developed a Hierarchical Probe Design (HPD) program that designs both sequence-specific probes and hierarchical cluster-specific probes from sequences of a conserved functional gene based on the clustering tree of the genes, specifically for analyses of functional gene diversity in environmental samples. HPD was tested on datasets for the nirS and pmoA genes. Our results showed that HPD generated more sequence-specific probes than several popular oligonucleotide design programs. With a combination of sequence-specific and cluster-specific probes, HPD generated a probe set covering all the sequences of each test set.

Availability: http://brcapp.kribb.re.kr/HPD/

Contact: yhpark@kribb.re.kr

Supplementary information: http://brcapp.kribb.re.kr/HPD/HPD_ Supplementary.doc

INTRODUCTION

Microorganisms in the environment have important roles in regulating the biogeochemistry of ecosystems. In order to evaluate the activities and functions of microbial communities, it is necessary to characterize the diversity of the functional microbial group of interest and to analyze individual members of the group under different environmental conditions. Because of the tremendous microbial diversity in the environment, extensive coverage is very important (Amann et al., 1995; Holben and Harris, 1995; Guschin et al., 1997; Curtis et al., 2002). To assess the structure and genetic potential of the microbial community, PCR-based molecular technologies, such as gene cloning, denaturing gradient gel electrophoresis and terminal restriction fragment length polymorphism have been used widely for assessing the structure and activities of the microbial community. In the case of carbon and nitrogen cycling genes, the pmox, nirS, nirK, nosZ and narG genes are highly conserved, and many of them could be retrieved directly from environmental samples (i.e. metagenomes) by PCR amplification using universal primers (Rosh et al., 2002; Dedhly et al., 2004).

To date, hundreds of gene sequences for a single functional gene have been deposited in the public database through this approach. Theoretically, these accumulated data resources make possible a comprehensive analysis of all these genes and their activities in the environments.

Recently, various types of DNA microarrays, such as cDNA microarrays and oligonucleotide microarrays, have been applied to study the microbial diversities of various environments and these arrays are useful because of their high throughput nature (Guschin et al., 1997; Wu et al., 2001; Taroncher-Oldenburg et al., 2003; Bodrossy et al., 2003). Previous research has shown that long oligonucleotide probes (50–70mer) have a better efficiency than short oligonucleotides (20–30mer) or cDNA probes in microarray-based diversity searches (Taroncher-Oldenburg et al., 2003; Tiobia et al., 2004; Rhee et al., 2004). The intensive labor in both PCR amplification and array fabrication limits the application of cDNA microarrays. In addition, it is a challenge to obtain all the diverse environmental clones and bacterial strains from various sources as templates for amplification. Moreover, the signal intensity from long oligonucleotide probes is several orders of magnitude above that of a short oligonucleotide microarray (Schröder et al., 2001, http://www.mwg-biotech.com/docs/discovery/an_arrays2_014.pdf).

Many computational programs have been developed for oligonucleotide probe selection, for instance, OligoWiz (Nielsen et al., 2003), PROBEmer (Emrich et al., 2003), Osprey (Gordon and Sensen, 2004), OligoPicker (Wang and Seed, 2003) and OligoArray 2.1 (Rouillard et al., 2003). These programs generate sequence-specific probes for each gene of a genome. However, if the target
sequences are highly similar to each other, truly sequence-specific probes cannot be generated because of cross-hybridization. In addition, it is difficult to define meaningful specific groups for probe design (Behr et al., 2000). To solve this problem, a cluster- or group-specific probe concept has been applied, specifically for highly conserved phylogenetic genes, such as 16S rRNA. The ARB probe design tool (Ludwig et al., 2004), PRIMROSE (Ashelford et al., 2002), Meier's analysis (2004) and Zhang’s analysis (2002) have proposed to design short oligonucleotide probes (C20 bp) from a group of sequences based on the 16S rRNA gene phylogenetic tree. The use of ARB has also been extended to the design of short oligonucleotides for the pmoA and amoA genes (Bodrossy et al., 2003). However, these programs provide insufficient parameters to design long oligonucleotides (>50 bp). The ARB uses a simple function of melting temperature as a parameter in probe design (ARB help, http://www2.mikro.biologie.tu-muenchen.de/arb/help_html/probe_param.html), which is sufficient to calculate melting temperature for short oligonucleotides, but insufficient for long oligonucleotides. For example, to get 50mer probes using ARB, the melting temperature boundaries need to be set to the unrealistic levels of 100–200°C.

To the best of our knowledge, there is no software package available for designing long oligonucleotide probes for highly conserved gene sequences. In this paper, we describe a new algorithm called Hierarchical Probe Design (HPD), which uses the concept of cluster-specific probes to cover target sequences in a cluster. HPD automatically generates probes against all nodes (clusters) of the clustering tree for sequences of a conserved functional gene. The scheme of Figure 1 shows how a hierarchical probe set designed by using this program helps us identify functional members of a microbial community.

**ALGORITHM**

**Definitions**

Before describing the procedure of designing hierarchical probes, the terminology and conventions used in this paper are briefly described. The probes generated by HPD include sequence-specific probes and hierarchical cluster-specific probes. A hierarchical cluster is a group of sequences that comprises a node in the clustering tree. The level of a cluster is determined by its hierarchical order. Based on the concept of hierarchical cluster, all input sequences are divided into ‘In-Cluster’ and ‘Out-Cluster’ groups according to their location in the tree. The level of a cluster is determined by its hierarchical order. The In-Cluster consists of the sequences within a hierarchical cluster, and the Out-Cluster are those sequences outside the cluster. A sequence-specific probe only hybridizes with a unique sequence, whereas a hierarchical cluster-specific probe may hybridize with all or a certain number of sequences in the In-Cluster, but it does not hybridize with any sequences in the Out-Cluster. Thus, the cluster-specific probes cover the sequences in a cluster completely or partially and are referred to here as full-cluster-specific probes and partial-cluster-specific probes, respectively. Because full-cluster-specific probes are the ideal solutions to discriminate a cluster, priority was given to making full-cluster-specific probes. For convenience,
an oligonucleotide probe is referred to as a probe and a hierarchical cluster as a cluster. Figure 3A shows an example of these categories of probes.

**HPD’s probe design process**

HPD has been developed to generate a probe set to identify the diversity of a highly conserved functional gene in complex samples. To achieve this goal, HPD is composed of three stages, as shown in Figure 2. First, the input sequences are aligned and then clustered. Second, probe candidates are generated for each cluster. Finally, optimal probes are selected based on experimentally validated criteria. Each step is detailed in the following sections.

**Multiple alignments and clustering.** To obtain a hierarchical cluster, HPD first incorporates ClustalW (Thompson et al., 1994) to conduct a multiple alignments for the input sequences. ClustalW compares each sequence successively to create a single alignment in which homologous residues are aligned in the same columns. Sequences containing gaps which shift the open reading frame were removed after the alignment was completed. User-generated aligned sequences can be imported into HPD to produce a user-oriented cluster or to accelerate the execution speed.

Based on these multiple alignments, sequences are clustered hierarchically by existing clustering methods. Our program extracts grouping information from the constructed tree to determine the hierarchical clusters. Here, for example, the neighbor-joining (Saitou and Nei, 1987) and UPGMA (Sokal and Michener, 1958) methods are utilized to cluster input sequences in HPD. However, to enhance the program’s flexibility, HPD has an optional function capable of importing a clustering tree generated externally. Any clustering methods can be employed, according to the user’s preference.

**Probe candidate selection.** The aim of this stage is to generate all possible probe candidates and assign each probe candidate to a cluster based on the probe specificity. As shown in Figure 3, the first step of the process starts by selecting probe candidates from each sequence. Next is the hierarchical process of cluster-specific probe candidate selection. When the process reaches the root cluster, the probe candidates residing in each sequence or cluster are assigned to be sequence- or cluster-specific probes. This process is explained in detail below.

**Step 1.** Probe candidate generation: the first step starts from the lowest level of a clustering tree as shown in Figure 3A. Each terminal is a cluster comprising one sequence (S1–S5). Initially, all possible probe candidates are generated against each of these input sequences, without any test of specificity with regard to the other sequences. Probe candidates are collected along with the sequences from 5' end to 3' end using a sliding window whose size is equal to the probe length. Each candidate has the possibility of being a sequence-specific probe or cluster-specific probe and selection of particular candidates proceeds according to the rules outlined below.

**Step 2.** Cluster-specific probe selection: this step checks the specificity of probe candidates against clusters one-level higher. Two sibling or hierarchically neighbored clusters are combined to generate a new upper cluster. For example, S1 and S2 form the C2 cluster, whereas S3 and S4 form the C4 cluster (Fig. 3B). The probe candidates generated at Step 1, against S1–S5 (Fig. 3A), are tested for their specificity toward these upper clusters (C2 and C4). If a candidate of one sibling cluster has enough specificity when tested against the sequence(s) of the other cluster to distinguish between these two clusters, it stays in the sibling cluster. If not, the candidate will be transferred to the next higher level of cluster. This process is repeated until all the sequences constitute one cluster, in this example C1. In other words, this process is terminated when the recursion reaches the root of the hierarchical cluster (Fig. 3D). After calculation of probe specificity, probe candidates specific to the C4 cluster, (e.g. Fig. 3B), can be selected as cluster-specific probes. All probe candidates that are not specific to the C4 cluster will be transferred to the associated upper cluster, i.e. C3 (Fig. 3C). In each step, probe specificity is checked by comparing a probe with the
associated target regions (at the same column of alignment) of the sequences in the In-Cluster and the Out-Cluster. This whole round of steps is repeated until it reaches the root cluster (C1 in Fig. 3D).

After reaching the stage of the root cluster, the probe candidates associated with clusters are judged to be cluster-specific probes. And the candidates associated with individual sequences only are judged to be sequence-specific probes.

Optimal probe set generation. The best probes set should show the maximum coverage of the associated In-Clusters and the maximum difference of identity between the In-Clusters and Out-Clusters. Additionally, all probes for each cluster are sorted based on several parameters of probe quality including cluster coverage, specificity, GC content and hairpin energy. The first sorting criterion applied is the region of sequence where a probe is designed. Since lengths of input sequences are different, there could be regions at 5' or 3' where alignment could not be made. The probes generated in these unaligned regions were excluded from the probe set. The second sorting criterion is the sequence coverage of In-Cluster sequences, which is a measure of the cluster adaptation of the probe. The third is the difference in identity between the target regions of the In-Cluster and Out-Cluster sequences, which sorts the probes in order to maximize specificity. Probes exceeding the predefined or program default limits of GC content or hairpin energy are removed from the sorted list. Finally, the specificity of the remaining probes is re-checked against the In-Cluster and Out-Cluster sequences using BLAST (Altschul et al., 1997). In the earlier stages of probe design (see above), HPD searches the sequences of the In-/Out-Cluster for possible cross hybridization only over the position where the probe was designed. Alignment to other regions of the sequence may allow cross hybridizations that generate false positives. Therefore, this BLAST filter step can assure the quality of the probe and screen out false positive probes. To choose the best hierarchical probe set from the pool of selected probes, each probe that has satisfied the above criteria is ranked by its specificity, i.e. fitness of finding associated clusters. Biophysical parameters of the selected probe, such as melting temperature, GC content, hairpin energy and hybridization free energy, are calculated and provided by HPD.

IMPLEMENTATION

The algorithm presented here was implemented in a computer program written in the Object Pascal programming language. The program incorporated ClustalW for multiple alignment and BLAST for local alignment. To calculate the folding energy of hairpin formation and the melting temperature of self-annealing, two programs were used, hybrid-ss-min and hybrid-min, which are included in the OligoArrayAux software package (Markham and Zuker, 2004, http://www.bioinfo.rpi.edu/applications/hybrid/OligoArrayAux.php). HPD was compiled with a Borland’s Delphi7 compiler.

HPD reads input sequences in the FASTA or GenBank format for alignment and also provides an interface for importing aligned sequences. This interface provides flexibility in editing aligned sequences or interfacing with other external programs, as the user prefers. HPD can load a clustering tree as well. The values for parameters, such as the oligonucleotide probe length, GC content boundary, identity threshold, hairpin energy limitation and \( T_m \) ranges, are options that can be modified by the user. The resulting probes are stored in a tabulated text file or in Microsoft Excel and...
the tabulated data for each probe include cluster ID, probe sequence, melting temperature, hairpin energy and hybridization free energy. Additionally, HPD provides information about the probe position in the sequence from which it was generated, the target sequences covered by the probes in the cluster and the cluster’s hierarchical position.

HPD requires Microsoft Windows NT, 2000 or XP as the operating systems. Its minimum requirement is Pentium 3 500 MHz with 256 MB memory. The size of memory required is dependant on the size of the target sequences. The computation time obviously depends on the number of input sequences. Here, we tested using 400 nirS gene sequences with 807.2 bp of average length. The correlation of computation time and size for sequences is described in detail in the Supplementary section.

Parameters of HPD

The default probe length was set to 50 bp based on the reports by Schröder et al. (2001) and Rhee et al. (2004). For hierarchical cluster-specific probes, the hybridization threshold of heteroduplex formation between the probe and the target sequences was set to −30 kcal/mol of hybridization free energy, 88% of probe-target identity and 15 bp of continuous stretch of matches on the basis of previous experimental data (Kane et al., 2000; Rhee et al., 2004; Steward et al., 2004). To remove the probes having hairpin secondary structures, the default for the self-folding energy threshold was set to −3 kcal/mol based on the reports by Bodrossy et al. (2003). To obtain the melting temperature, free energy rules were applied at 65°C. The GC content ranged between 35 and 65%. The number of final probes for each cluster was limited to one, unless otherwise specified.

RESULTS AND DISCUSSION

To evaluate HPD, we selected two ecologically important genes involved in the nitrogen and carbon cycles: nitrite reductase (nirS) and methane monooxygenase (pmoA). The sequences downloaded from GenBank (http://www.ncbi.nlm.nih.gov) in June 2004 varied in length (128–1791 bp). Most of the deposited sequences were partial and extracted directly from environmental samples. To generate a high-quality sequence alignment to use in HPD, the nirS sequences were filtered by a threshold value of 700 bp minimum, and the pmoA sequences were filtered by a threshold value of 450 bp minimum. The resultant test sets contained 421 nirS sequences and 490 pmoA sequences. All the sequences were conserved. The nirS sequences had an average identity of 68.4 ± 7.5% (standard deviation) and the pmoA sequences had an average identity of 69.5 ± 6.4%.

Probe generation from nirS and pmoA test sets

Using the default constraints described in the Implementation section, a total of 380 possible probes (145 sequence-specific probes and 235 cluster-specific probes) were found in the nirS set, and 234 probes (63 sequence-specific and 171 cluster-specific) were found in the pmoA set (Table 1). Sequence-specific probes covered 35% of nirS sequences and 13% of pmoA sequences, respectively. The number of pmoA sequence-specific probes is less than that of nirS because of the higher pairwise sequence similarity in pmoA than in nirS (data not shown). When we included the hierarchical cluster-specific probes, the probe sets covered all the sequences for both nirS and pmoA test sets. For the clusters in which full-cluster-specific probes could not be designed, partial-cluster-specific probes were selected, as described in the Algorithm section. In nirS and pmoA, the number of full-cluster-specific probes was larger than the number of partial-cluster-specific probes (Table 1). The full-cluster-specific probes for nirS covered 304 sequences, ~72% of the test set. In the case of pmoA, 479 sequences were covered by full-cluster-specific probes, ~98%. However, the nirS full-cluster-specific probes were localized to the lower levels of the hierarchy, and the sequence coverage per probe was much smaller. In addition, the nirS test set had fewer clusters which contained sequences with high identities (>90%) to each other at the bottom of the tree than pmoA. Thus, full-cluster-specific probes covered more sequences in pmoA than nirS.

Effect of clustering tree on hierarchical probe design

HPD requires a rooted tree as an input, because it selects probe candidates in a hierarchical manner as described in Algorithm section. Since probe design is dependent on the clustering method, we examined the effect of the clustering method and tree topology in generating an optimized probe set. In the present study, we compared the trees clustered by UPGMA and neighbor-joining methods. UPGMA is a simple clustering method by which local topological relationships are identified in the order of similarity. Neighbor-joining constructs a tree by linking the least-distant pair of nodes with a different kind of cluster results. Since the neighbor-joining method generates an unrooted tree, the sequence showing the longest distance in the UPGMA tree was selected as a default root. Although Maximum parsimony is a site-dependent clustering method, it could be used successfully for probe generation. The neighbor-joining, UPGMA and Maximum parsimony trees generated a similar number of probes for both nirS and pmoA (Table 1). By using the neighbor-joining clustering method of ClustalW, a total of 380 probes were found for nirS and 244 for pmoA. When we compared the cluster identity between the UPGMA and neighbor-joining trees, 279 clusters (57%) were identical in pmoA and 227
clusters (54%) were identical in nirS. In the identical clusters, HPD generated the same cluster-specific probes (data not shown). This suggests that HPD works consistently on different clustering trees.

Impact of identity threshold and probe length on hierarchical probe design

To evaluate the impact of the identity threshold on probe design, we tested identity thresholds ranging from 84 to 92% (Fig. 4). Here, the sequence coverage concerns only the In-Cluster sequences since probes generated show specificity only to sequences of the cluster. When all input sequences are assumed as 100%, ‘cumulative’ means sum of the unique (not redundant) sequences covered by the cluster-specific probe and subcluster-specific probes. Our results demonstrated that the cumulative sequence coverage decreased in small-sized clusters (i.e. clusters close to the terminal nodes), as identity threshold decreased. With lower identity thresholds, although the number of probe candidates increased, more probe candidates had the potential to cross-hybridize to sequences of the Out-Cluster. In nirS, we observed a relatively smaller decrease in cumulative sequence coverage attributed to changes in the threshold value, than in pmoA. Since nirS is more diverse in sequences, less effect would be expected on the number of cluster-specific probes by changes in the identity threshold.

To evaluate the effect of probe length on probe design efficiency, we varied the probe length from 20 to 70 nt (Fig. 5). Our results indicate that a probe set 20 bp in length showed greater coverage for a set of small-sized clusters than probe sets of 40–70 bp in length. Changes in probe length in the range of 40–70 bp did not show any significant effect on the coverage of the probe set. Zhang et al. (2002) reported that signature sequences (in the range of 5–15 bp) have a polar distribution in the tree; i.e. longer oligonucleotides tend to identify the clusters near the leaves (small-sized clusters), whereas short oligonucleotides are more likely to pick out clusters near the root (large sized clusters). However, this effect was not observed in probe sets with a length >20 bp in HPD.

Fig. 4. Impact of identity threshold on probe designs (A) nirS and (B) pmoA. The number of sequences covered by probes increases as the size of the cluster increases. Each probe is a 50mer oligonucleotide covering a sequence or a cluster of sequences. Probe sets were generated using 84, 88 and 92% identity thresholds as default parameters.

Fig. 5. Impact of probe length on probe designs (A) nirS and (B) pmoA. The number of sequences covered by probes increases as the size of the cluster increases. Each probe is a 50mer oligonucleotide covering a sequence or a cluster of sequences. Probe sets were obtained by using probe lengths of 20, 40, 50 and 70 bp as default parameters.
Assessment of HPD using a small sample set of sequences

The results of HPD are displayed in detail in Figure 6 for two small sets of sequences. A sample set that contains 47 \textit{nirS} sequences having 64\% average identity was selected from a local region of the tree constructed from 421 \textit{nirS} sequences. Another sample set that contains 50 \textit{pmoA} sequences having 85\% average identity was picked from a local region of the tree constructed from 490 \textit{pmoA} sequences. The quality and distribution of the probes is shown by combining the sequence-specific and cluster-specific probes generated using the default parameters (Fig. 6). The sequence identity of a cluster in the \textit{nirS} sample set is rather lower than that in the \textit{pmoA} sample. The \textit{nirS} sample set has 9 subgroups and 24 sequences within the 95\% identity threshold (see the line \(x\) of Fig. 6). In contrast, the \textit{pmoA} sample set has 11 subgroups and 2 sequences within the 95\% identity threshold. As a result, 24 sequence-specific probes were found for the \textit{nirS} sample set among 55 probes, whereas only two sequence-specific probes were found for the \textit{pmoA} sample set among 27 probes. The sequence-specific probes were designed for sequences showing <95\% identity to the other sequences. The full-cluster-specific probes of small clusters showed a relatively high specificity (more black squares as shown in Fig. 6). A cluster-specific probe which is close to the root cluster (e.g. 40 and 42 in Fig. 6A; 18 and 26 in Fig. 6B), has some gray squares (<100\% identity to target) in the cluster sequences. However, the full-cluster-specific probes close to the terminal clusters have a 100\% probe-target identity. Two clusters (P23 and P24 in Fig. 6B) in the \textit{pmoA} sample set did not have any candidate probes covering all the In-Cluster sequences, and, therefore, partial-cluster-specific probes were selected. Twelve clusters in the \textit{nirS} sample set (41, 44 and 46–55 in Fig. 6A) did not have any candidate probes covering all the In-Cluster sequences; thus, partial-cluster-specific probes were selected for these clusters. The missing specificity of a cluster could be complemented by its subcluster (not other partial cluster) specific probes. For example, in the Figure 6A, partial-cluster probe 43 could be complemented by subcluster’s probes, such as 5, 6, 25, 26 and 27.

Comparison of HPD with other probe design programs

To assess the performance of the HPD software, we compared HPD with two other probe design software packages, OligoPicker and OligoArray 2.1, using three sample datasets. Since there are no available software packages that generate long oligonucleotide cluster-specific probes from functional genes, we could only compare the ability to generate sequence-specific probes. We used three sample sets for comparison: the two sample sets used for testing the performance of HPD above and the first 44 ORF sequences from the \textit{Escherichia coli} K-12 genome (ftp://ftp.ncbi.nih.gov/genbank/ genomes/Bacteria/Escherichia\_coli\_K12/). HPD utilized the default parameters described earlier. The parameters of OligoPicker and OligoArray 2.1 were set to the program defaults, with the
false positives. HPD generated 24 sequence-specific probes and shown to hybridize other non-target sequences, thus generating probes were sequence specific, and the other 33 probes were generated 47 probes for the results from OligoArray 2.1. As shown in Table 2, OligoArray 2.1 results for these datasets (data not shown), we only present the

CONCLUSION

In summary, we present here a new probe design software called HPD to generate sequence-specific and cluster-specific long oligonucleotide probes for sequences of a conserved functional gene. Owing to the high degree of similarity between the sequences within each functional gene category, it has been impossible to design sequence-specific probes for all input sequences. HPD attempted to find hierarchical cluster-specific probes, which cover the target sequences of a cluster within a hierarchy.

HPD is capable of producing multiple probes per sequence or cluster to cover an input sequence set in parallel and in hierarchy. When multiple probes of different hierarchical levels are combined for diversity analysis of a functional gene, the confidence in identifying target microorganisms from the microbial community will be enhanced. In the future, we will focus on increasing the probe design efficiency by incorporating faster alignment tools and developing a heuristic search algorithm that considers hierarchical clustering.

ACKNOWLEDGEMENTS

We thank Sam-Myo Kim and Jong Bhak for helpful discussion on the algorithm. This work was supported by grants M1040701000-10SN070100100, NNM0100411, BDM0200413; NRL research programme (M10104000294-01000012800) and 21C Frontier Microbial Genomics and Application Center Program (MG05-0104-3-04) from the Ministry of Science and Technology (MOST), and the KIRIBB Research Initiative Program, Republic of Korea.

Conflict of Interest: none declared.

REFERENCES


