Dual-genome primer design for construction of DNA microarrays

Anders Andersson¹,², Rolf Bernander² and Peter Nilsson¹,*

¹Department of Biotechnology, KTH—Royal Institute of Technology, SE-106 91 Stockholm, Sweden and ²Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, SE-752 36 Uppsala, Sweden

Received on June 29, 2004; revised on August 19, 2004; accepted on August 20, 2004
Advance Access publication August 27, 2004

ABSTRACT

Motivation: Microarray experiments using probes covering a whole transcriptome are expensive to initiate, and a major part of the costs derives from synthesizing gene-specific PCR primers or hybridization probes. The high costs may force researchers to limit their studies to a single organism, although comparing gene expression in different species would yield valuable information.

Results: We have developed a method, implemented in the software DualPrime, that reduces the number of primers required to amplify the genes of two different genomes. The software identifies regions of high sequence similarity, and from these regions selects PCR primers shared between the genomes, such that either one or, preferentially, both primers in a given PCR can be used for amplification from both genomes. To assure high microarray probe specificity, the software selects primer pairs that generate products of low sequence similarity to other genes within the same genome. We used the software to design PCR primers for 2182 and 1960 genes from the hyperthermophilic archaea Sulfolobus solfataricus and Sulfolobus acidocaldarius, respectively. Primer pairs were shared among 705 pairs of genes, and single primers were shared among 1184 pairs of genes, resulting in a saving of 31% compared to using only unique primers. We also present an alternative primer design method, in which each gene shares primers with two different genes of the other genome, enabling further savings.

Availability: The software is freely available at http://www.biotech.kth.se/molbio/microarray/
Contact: anders.andersson@biotech.kth.se

INTRODUCTION

DNA microarray technology has become a widely used molecular biology tool for determining transcriptional states of cell populations. Even though the number of completely sequenced free-living species now exceeds 150, the majority of microarray studies published to date have been performed on a limited number of model organisms. The field of comparative genomics has proven the power of comparing sequences of species at different evolutionary distances for predicting gene functions and identifying regulatory elements (Tatusov et al., 1997; Overbeek et al., 1999; Pellegrini et al., 1999; Wasserman et al., 2000). In contrast, transcript profiling studies have so far focused on expression patterns of single organisms. As the amount of expression data from a variety of species increases, this opens the door for cross-species comparisons (Teichmann and Babu, 2002; Alter et al., 2003; Stuart et al., 2003; van Noort et al., 2003; Bergmann et al., 2004). This will be important for identifying functionally related genes; if a set of genes displays similar expression patterns not only in one, but in several species, the probability that the genes are functionally related, rather than co-expressed by chance, increases. Orthologous genes displaying similar expression patterns are likely to contain conserved regulatory sequences, which subsequently can be mined. Comparative studies are also important for separating species-specific features from those that are more generally conserved.

A major drawback of the microarray technology is high production costs. Even though prices for synthetic oligonucleotides have decreased significantly during the last years, costs for producing whole-genome arrays are considerable, as each gene should be represented by a gene-specific oligonucleotide probe, or by a PCR product generated with a pair of gene-specific primers. By using cDNA libraries for the production, the oligonucleotide costs can be reduced, but instead costs for cloning and sequencing are introduced, and reduced probe specificity is often obtained. The high costs often force researchers to limit their studies to a single species, even though comparison of gene expression patterns between several related organisms would greatly enhance the information gain.

Computational methods for finding reduced sets of primers sufficient to amplify all genes in a genome have been proposed (Fernandes and Skiena, 2002). These rely on short primers, each occurring in multiple genes, which are combined such
that the primers of a selected pair occur close to each other, in the correct orientations, in a single gene. Even though it is computationally feasible to find such primer sets, these short primers have not been shown to work adequately in actual PCRs.

When designing probes for several genomes, advantage can be taken of sequence similarities between the different genomes. Methods have been developed for designing common primers for multiple, aligned, sequences (Gibbs et al., 1998). This is a straightforward strategy when designing primers for a few genes with known similarity, but not suitable for design at the genome level. We have developed a method, implemented in the software DualPrime, which utilizes small common regions between two genomes, not necessarily within homologous genes, and designs primers within these. Preferentially, both primers should be shared, and as many pairs of genes as possible sharing primer pairs are identified. Among the remaining genes, as many pairs of genes as possible that share single primers are identified. The actual primer design is performed by the Primer3 software (Rozen and Skaletsky, 1998, http://www.genome.wi.mit.edu/genome_software/other/primer3.html), which is fed with gene sequences and coordinates for where to perform primer searches.

**METHODS**

**Outline of method**

The software works as follows (Fig. 1): short regions of high sequence similarity are identified between the two genomes. For each gene that shares at least two regions with a gene in the other genome, primers are designed within the shared regions. Among the primed genes, as many pairs of non-redundant genes as possible are selected that share primer pair.

For the remaining genes, new primers are designed, this time requiring at least one of the primers to be located within a region shared with another gene. Among the genes primed in this round, as many pairs of non-redundant genes as possible are selected that share one primer. Finally, the remaining genes are primed with unique primers. The actual primer design is performed by the Primer3 software (Rozen and Skaletsky, 1998, http://www.genome.wi.mit.edu/genome_software/other/primer3.html), which is fed with gene sequences and coordinates for where to perform primer searches.

![Fig. 1. Outline of dual-genome primer design by DualPrime.](image-url)
To avoid cross-hybridizations on the microarrays, the PCR products need to be sufficiently specific for their target genes. The software assures this by only accepting primers that do not generate amplicons with 67% identity or above, over more than 30 nt, to another gene than the intended, in the same genome.

Finding similar regions
To enable robust PCR amplifications, the length of the primers was set to ≥18 nt. Therefore, this is the minimum length of the similar regions searched for. We have found that the presence of two degenerate positions within a primer does not significantly alter the result of the PCR, provided that none of the positions is located at the 3′ end (data not shown). Thus, to increase the number of similar regions, we allow two mismatches within the 18 nt.

For each gene, all degenerate and perfectly matched 18mers that this gene shares, in either forward or reverse complementary direction, with any gene in the other genome (called partner gene), are found by an algorithm specifically tailored to identify 18mers with a maximum of two mismatches. If several 18mers shared with a specific partner gene in a specific direction overlap, or are adjacent, these are merged into longer shared regions. To avoid unspecified priming, all 18mers occurring multiple times within a genome are discarded.

Designing primers
The primer design is performed by the Primer3 software. Our software uses the input parameters ‘excluded_region’ (primers are not allowed within this region) and ‘target’ (the primer pair must flank this region) to force Primer3 to locate primers within regions of interest.

Parameters regarding primer quality (minimal primer size, GC content, self complementarity, etc.) are either set by the user or default values are used. Output primer pairs containing a primer with an N at the 3′ end are discarded, as well as primer pairs generating unspecified amplicons (see below).

Three different categories of primers are designed during the process:

1. Both primers are located within regions shared with a partner gene. For each direction (forward and reverse complementary), and for each partner gene that the gene shares more than one region with, an input record (file with input parameters to Primer3) is created for every pair-wise combination of shared regions. Regions outside the two shared regions are marked as ‘excluded_region’. Any mismatched nucleotide within the shared regions is replaced with an N in the input sequence. A maximum of 50 output primer pairs are generated for each input record.

2. At least one primer is located within a region shared with a partner gene. For each partner gene, and for each region shared with the partner gene, two input records are created. To assure that one of the primers is located within the shared region, all sequence upstream of the region is, in the first input record, marked as ‘excluded_region’, and the region immediately downstream of the first nucleotide is marked as ‘target’. The opposite is done for the second input record, to generate primers reverse complementary to the shared region. If a mismatched nucleotide occurs within the shared region, this is replaced by an N in the input sequence. A maximum of 50 output primer pairs are generated for each input record.

3. Primers are allowed anywhere in the sequence. A maximum of 50 output primer pairs are generated for each gene.

Specificity control of PCR products
After each category of primers is designed, the specificity of the amplicons to their target genes is tested. A similarity search is performed with BLAST (Altschul et al., 1990). Instead of searching with each amplicon (as this would be very time consuming), each gene is searched once against all other genes within the same genome. By setting the mismatch penalty to −2, the chances of finding longer regions of similarity increase. The positions of all regions within the gene giving non-self hits >30 nt in length, with 67% identity or above, are stored in a hit table. The primer pairs are then compared with the hit table, and those generating amplicons with >30 nt overlap with a non-self hit are discarded.

Finding pairs of genes sharing primer pairs
The primary goal is to find as many gene pairs as possible sharing primer pairs. The software attempts to design category 1 primers (above) for as many genes as possible. With the output primers that passed the specificity control, a hash table is created of all primer pairs used for the first genome, ‘{primer1}{primer2} = gene name’, and this is searched with output primer pairs from the second genome. As each mismatched position within shared regions has been replaced by an N in the input sequences, output primers for a pair of partner genes, located within shared regions will, despite mismatches, have identical sequences, and may be compared simply by exact string matching. If a primer pair was used for both genomes, the gene pair sharing the primer pair is saved in a new hash table, the gene pair table: ‘{gene_name1}{gene_name2}’.

As a gene can share primer pairs with several genes, the task is now to select as large a set of gene pairs as possible, in which each gene occurs only once. To do this we build a bipartite graph, where the nodes of the first and the second partition represent genes of genome 1 and genome 2, respectively (Fig. 2). If a pair of genes share primer pair/pairs, their corresponding nodes will be connected by an edge in the graph. To maximize the matching in the graph,
that is, select as many edges as possible with unique nodes, an algorithm based on the Ford–Fulkerson method. (Ford and Fulkerson, 1962) for solving the maximum-flow problem is used.

Most of the selected gene pairs share several primer pairs. We have observed that signal intensity on the microarray is positively correlated with the length of the PCR product (data not shown), and therefore we want to avoid unnecessary short PCR products. Thus, for each primer pair, the shorter of the two amplicons is identified, and for each gene pair the primer pair with the longest (of the shorter) amplicon is selected. Finally, any N that may exist in the primer sequences is replaced by mixed nucleotides, based on the original gene sequences, according to the IUB (International Union of Biochemistry) code.

Finding pairs of genes sharing single primers

Among the remaining genes, we want to find as many pairs of genes as possible sharing single primers. The program designs category 2 primers for as many of these genes as possible. We use the output primers to create a new hash table of all primers used for genome 1: ‘{primer1} = gene-name’, ‘{primer2} = gene name’, and search this with all primers used for genome 2. All pairs of genes sharing a primer are stored in a new gene-pair hash table ‘{gene_name1}{gene_name2}’, and as large a set of gene pairs of non-redundant genes as possible is selected in this table using the same method as above. For each selected gene pair, we choose the primer pairs (one shared and two genome-specific primers), with the longest amplicons. Each N in the shared primer sequences is replaced by mixed nucleotides according to the IUB-code.

Implementation

The DualPrime software was implemented in Perl (version 5) and makes use of the publicly available programs Primer3 and BLAST.

Designing primers for *S. solfataricus* and *S. acidocaldarius*

We allowed amplicons in the size range of 100–1000 bp. Primer lengths of 18–22 nt were permitted, with melting temperatures between 48 and 65°C. The maximum difference between the melting temperatures in a pair of primers was set to 6°C, and the maximum length of a mononucleotide repeat in a primer was set to 4. Default values were used for all other Primer3 parameters.

PCR amplifications and fabrication of microarrays

To overcome problems with melting temperature differences between the primer pairs, we used a touchdown PCR protocol (Don et al., 1991), in which the annealing temperature was lowered by 0.5°C per cycle, during the first cycles (from 50 to 45°C). PCR amplifications were performed in 100 μl reaction volumes. All PCR products were inspected on agarose gels before purification with MultiScreen SEQ384 filter plates (Millipore). Purified PCR products were dissolved in 60 μl of 50% DMSO and printed in triplicate on Ultra GAPS microarray slides (Corning) using a QArray instrument (Genetix).

Microarray hybridization, scanning and analysis

Labeled genomic DNA and cDNA were prepared and hybridized as described at www.biotech.kth.se/molbio/microarray. Microarray slides were scanned with an Agilent Microarray Scanner (Agilent Technologies). Image processing was performed using the GenePix Pro 5.0 software (Axon Instruments). Low-quality spots were excluded by filtering out those for which the ratio of medians deviated from the regression ratio by more than 20% or if, for any channel, <70% of the foreground pixels had intensities exceeding median background plus 2SD of background.

RESULTS

We applied DualPrime for the construction of microarrays for the two hyperthermophilic archaea *S. solfataricus* and *S. acidocaldarius*. The 2.99 Mb genome of *S. solfataricus* encodes 2997 (She et al., 2001) and the 2.22 Mb genome of *S. acidocaldarius* approximately 2268 proteins (L. Chen, personal communication), respectively. The primer design was restricted to protein-coding genes longer than 300 bp. After removal of transposases, 1991 *S. acidocaldarius* genes and 2438 *S. solfataricus* genes remained.

Finding similar regions

Assuming random distribution of the four nucleotides, a given 18mer would occur approximately $4.4 \times 10^{-5}$ times
in the 2.99 Mb *S. solfataricus* genome. By allowing up to two mismatches, the frequency increases 1432-fold \([C(18, 2) \times 3^2 + 18 \times 3 + 1]\), and an *S. solfataricus* gene would share at least one degenerate 18mer with on average 150 genes in *S. acidocaldarius*. The corresponding number in *S. acidocaldarius* was 183, the larger number reflecting the larger number of *S. solfataricus* genes. All of the 2438 *S. solfataricus* genes shared at least one degenerate 18mer with at least one of the 1991 *S. acidocaldarius* genes, and vice versa.

### Selection of primers

*S. solfataricus* contained 2003, and *S. acidocaldarius* 1684, genes that shared at least two regions with at least one partner gene in the other genome. Among these, the software rendered output primers located in the shared regions (category 1 primers) for 977 *S. solfataricus* and 961 *S. acidocaldarius* genes. After filtering out primer pairs generating unspecific amplicons, 860 and 920 genes remained, respectively. From this set, the software found 705 gene pairs of non-redundant genes sharing primer pairs.

Among the remaining genes, 1729 of the 1733 *S. solfataricus* and all of the 1286 *S. acidocaldarius* genes generated output primers belonging to category 2, i.e. one primer located within a region shared with a gene in the other genome. After filtering out primer pairs rendering unspecific amplicons, 1539 and 1266 genes remained, respectively. Among these, 1439 genes in *S. acidocaldarius*, and 1211 genes in *S. solfataricus*, shared a primer with at least one gene in the other genome. In total, 6564 gene pairs were formed (each gene in *S. acidocaldarius* shared at least one primer with on average five *S. solfataricus* genes). Among these, the program found 1184 gene pairs of non-redundant genes.

The remaining genes had to be primed with unique primers from category 3. However, the software failed in finding primer pairs with sufficiently gene-specific amplicons for 31 of the 102 *S. acidocaldarius* and 256 of the 549 *S. solfataricus*, genes. Sequence analysis revealed that the majority of these were homologous to one or several other failed genes, indicating that they belonged to gene families. This is in agreement with earlier estimations of the number of gene families within the *S. solfataricus* genome (She et al., 2001).

Table 1 summarizes the number of primers used. In total, 5690 primers were selected to prime the 4142 genes (otherwise requiring 8284 unique primers), which resulted in a saving of 2594 primers (31%).

### PCR and microarray hybridization results

When inspected on agarose gels, \(~98%\) of the PCRs gave products with single bands of expected sizes, whereas 0.3% produced double bands, and the remaining were empty (Fig. 3). After purification, the PCR products were printed on microarray slides. These were hybridized with Cy3- and Cy5-labeled genomic DNA from *S. acidocaldarius*. For 97% of the PCR products that displayed single bands on the gels, two out of three spots passed quality filtering, rendering a final success rate of 95%. Of the probes that passed the quality filtering, 95% gave ratios with <10% deviation from the array average ratio (Fig. 4a). To further test the accuracy of
to avoid cross-hybridizations within the genomes (the average global alignment homology between the S.solfataricus probes and their most similar S.acidocaldarius genes was 72%). None of the probes with similarities below the cut-off level (55% of the S.solfataricus probes) gave cross-hybridization signals. The result from this cross-species hybridization is a good indication that the requirements for specificity demanded by our software are sufficient to avoid cross-hybridization within the genomes.

**The primer path approach**

With the approach described above, the number of primers needed for two genomes can, in the ideal case, be reduced by 50%. This requires equal numbers of genes in the two genomes, and that each gene shares at least two regions, suitable for primer design, with a gene in the other genome. This means that the species must be closely related, as it is unlikely that two regions, separated by an appropriate distance and with similar GC content etc., are shared between a pair of genes by chance. However, a single region is often shared even between unrelated genes, and most genes share regions with many genes (above). To utilize this fact, we propose a further development of the method, where each gene shares its forward and reverse primer with two different genes of the other genome. Thus, for each gene, primers are designed within each pair-wise combination of regions shared with different genes.

The genes and output primers generated in this way may be represented by a graph (Fig. 5), in which primers are represented by nodes and genes by edges. A primer is represented by a single node, whereas a gene may be represented by several edges (if primed by several primer pairs). If the edges are given gene-specific colors, we wish to choose as small a set of nodes as possible that connects edges of all different colors. Thus, a set of paths will be selected in the graph, and in each path the number of primers will be equal to one plus the number of primed genes within the path. Inspired by Fernandes and Skiena (2002) heuristic for finding a minimal primer set, we use a greedy algorithm performing the following operations:

1. For each edge color in the graph, calculate the 'rareness', i.e. 1 divided by the frequency of the color.
2. For each node, compute a weight as the sum of the rareness of all adjacent edges. That is, a node inducing many edges with colors of low frequency will get a high weight.
3. For each color, select an edge with that color as seed edge. Select the edge with the maximum sum of weights.
4. Randomly select an edge in the graph. Test whether replacing the seed edge of the same color with this will decrease the number of nodes in the subgraph. If so,

**Cross-hybridization**

Of the 705 gene pairs that were amplified with common primer pairs, 595 yielded PCR products of similar lengths (±3 bp) in the two genomes, indicating that regions from homologous genes were amplified (Fig. 3). Still, only 10 of these probes gave cross-hybridization between the genomes, when labeled genomic DNA from S.acidocaldarius was hybridized on dual-genome arrays. These 10 probes were all found to display similarities exceeding the cut-off levels used by our software. 

![Graph](image-url)
change the seed edge. If the number of nodes remains unchanged, randomly, with equal probability, either keep the old or replace the seed edge.

(5) Repeat Step 4 until no further decrease in the number of nodes in the subgraph is obtained.

Using this algorithm, 5019 primers were identified that together prime the 4142 genes of \textit{S. solfataricus} and \textit{S. acidocaldarius}. This means a reduction with another 671 primers, as compared to the first method described. A drawback with this approach is the logistic problems arising when the primers are to be combined for PCR. However, by positioning the primers in an organized manner in 96-well plates, these problems can be circumvented; in the first 96-well plate the first primers of 96 paths of similar lengths, all starting with (for instance) genome 1, are placed. In the second plate, the second primers of the paths are placed. In the third, the third primers, etc. To run the PCRs, plate one and two are combined for PCR. However, by positioning the primers in an organized manner in 96-well plates, these problems can be circumvented; in the first 96-well plate the first primers of 96 paths of similar lengths, all starting with (for instance) genome 1, are placed. In the second plate, the second primers of the paths are placed. In the third, the third primers, etc. To run the PCRs, plate one and two are combined to prime 96 genes of genome 1, two and three to prime 96 genes of genome 2, etc.

**DISCUSSION**

We present two strategies for reducing the number of primers needed for the construction of dual-genome microarrays. We demonstrate the use of the first strategy for the production of whole-genome microarrays for two archaean species, and prove the successful use of the designed primer set for genomic-scale PCR amplifications. The resulting high-array quality is shown here and this has also been utilized in another study (Lundgren et al., 2004). The strategy is an attractive alternative when resources are limited, offering the opportunity to amplify a substantial fraction of the genes of a second, closely related, species of interest without increasing the primer costs. The second approach is useful for dramatically reducing the number of primers, and thereby the production costs, when microarrays are to be developed for multiple genomes, regardless of the evolutionary relatedness of the corresponding organisms.

**ACKNOWLEDGEMENTS**

We thank Jasper Walter, Annalie Waldén and Angela Fredin for technical assistance, Kim Brugger, Magnus Rosenlund and Lars Arvestad for valuable discussions, Per Unneberg for critical reading of the manuscript, and Lamming Chen and Roger Garrett for kindly providing gene sequences for \textit{S.acidocaldarius}. This work was supported by the European Union 5th Framework Programme.

**REFERENCES**


