AIMS Bioengineering

AIMS Bioengineering, 4(1): 28-45.
DOI: 10.3934/bioeng.2017.1.28
Received: 12 December 2016
Accepted: 26 December 2016
Published: 04 January 2017

http://www.aimspress.com/journal/Bioengineering

Research article

Sequence data analysis and preprocessing for oligo probe design in microbial genomes

Ruming Li [1],*, Brian Fristensky [1], and Guixue Wang [2],*

[1] Bioinformatics Laboratory, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
[2] Key Laboratory for Biorheological Science and Technology of Ministry of Education, Bioengineering College of Chongqing University, Chongqing 400044, China

* **Correspondence:** Email: rli@alumni.lsu.edu, wanggx@cqu.edu.cn; Tel: +1-604-417-1789, +86-236-511-2675.

Abstract: A good oligo probe design in DNA microarray experiments is crucial to obtain the better results of gene expression analysis. However, sequence data from a very large microbial genome or pan-genome will produce a reduced number of oligos and affect the design quality if processed by a probe designer. Gene redundancies and discrepancies across resources of the same species or strain and their sequence similarity and homology are responsible for the poor quantity of oligos designed. We addressed these issues and problems with sequences and introduced the concept of open reading frame (ORF) sequence segmentation from which quality oligos can be selected. Analysis and pre-processing of sequence data were performed using our Perl-based pipeline ORF-Purger 2.0. ORFs were purged of redundancy, discrepancy, invalidity, overlapping, similarity and, optionally, homology, such that the quantity and quality of oligos to be designed were drastically improved. Probe integrity was proposed as the first probe selection criterion since the fully physical availability of all possible probes corresponding to their targets in a nucleic acid sample is necessary for a best probe design.

Keywords: sequence data analysis; gene sequence pre-processing; ORF purging; ORF segmentation; DNA microarray; oligo probe design; oligo probe quantity; oligo probe quality

1. Introduction

DNA microarray has played a central role in present-day high-throughput gene expression analysis and other genomic studies since its advent. Among two types of DNA microarrays (cDNA or PCR and oligo), oligo-based microarray has received the most attention for many years and has been employed in most microarray experiments [1–6]. This popularity of oligo-based DNA microarray has been attributed to the rapid advance of genomic sequencing and oligo synthetic technologies. One of the important attributes for a successful microarray experiment is primarily contingent upon its quality of microarray design. A desired microarray design in turn depends largely on the quality of probe design [1,4,6–12]. With high-density DNA chips, it allows much more probes spotted on a slide but brings a new challenge as well in oligo design due to its sheer quantity. Since genomic information grows so fast with more genomes being sequenced that a great volume of sequence data have become a big issue today to tackle. In recent years, more and more interrelated microbial genomes with closer genetic makeup and sequence homology have been found and a bunch of gene families with highly similar transcripts and metabolites can be discovered accross microbial strains and even homologous species. This produces sequence data not only in huge amounts but also of the substancial similarities among them, which makes oligo probe designs much more difficult than ever [13–18].The sources of growing sequence data and increasingly similar nt or base compositions stem in part from many databases with different data acquisition and curation as well as various sequence submissions [16,19,20]. And they stem in large part from discrepancies introduced from the processes of evolutionary convergences or divergences, cumulative mutations, and the levels or depths of human working on genomes.

In addition to sequence data submitters such as numerous labs with differential data generating capabilities, different database curators and data management strategies contribute to the diversity and complexity of same or similar genomic sequences. For example, the sequence data from NCBI, JGI, EMBL, and the like have different types of FASTA-like file or have different formats of the same type, such as varying components and/or their arrangements in the header line of a FASTA file. They also differ in the gene predicting tools as well as the type and number of contigs used for genome annotation, which results in diverse yet redundant sequence data or discrepant gene entries of the same genome across data sources. Because almost all sequence data available are currently gathered from as many as databases or pooled with data of user's own for an inclusive collection of genes of interest, these overall genomic sequences or a superset of genes are informative for an inclusive oligo probe set. With the costly operation of microarray experiments, however, it holds up oligo designs of a particular genome and instead turns to cost-effective pangenome-based design thanks to accommodable room of spots on a high-density DNA chip. Other possible non-financial reasons for exploiting high-density biochips are for a joint or comparative study on multiple related strains or mutants and for a comparative genomic hybridization [18,21]. As a result, a broad range of genomic sequences enrichs the gene pool for a probe search with no or the least loss to necessary information for constructing an inclusive probe set. The inclusiveness or completeness of oligos spanning a whole genome is of great genetic significance to warranting that all possible target sequences have their respective hybridization probes. Presence of a full correspondence of each target with its probe is of great molecular significance to minimizing probe-nontarget cross-hybridization. The rationale underlying this design is that availability of direct, strong and stable probe-target binding would definitely preclude or anticipate probe-nontarget binding by preempting the hybridization reaction. Thus the attempted integrity of probes for a genome-wide or pan-genome coverage would be critical to not only global gene expression profiling and new gene discovery but also gaining accurate hybridization signals. The property of pangenomic data and the integrity of their probes are particularly crucial to very large microbial genomes (> 5,000 genes), as they would pose a higher risk of potential cross-hybridization and be more difficult in probe design. Because of these high-volume pangenomic data, they need to be reasonably classified and analyzed, cleaned, purged, and integrated before proper oligos can be constructed from them. For instance, with even the same type or format of sequence data, there may be some of differences among files in the number of nt bases (e.g., 50, 60, 70, 80) per line and the sporadic lowercased base letters (all bases must be uppercased) might be hidden problems for a program not taking care of them. Other aspects in structuring a data file (e.g., a greater sign > missing and spacing around it from the FASTA header lines) could also be pitfalls for a non-robust handling algorithm and eventually lead to an unsound result of data processing. Therefore, all of this needs to be tackled to ensure a reliability of output information. Besides these issues and snags, the prevalently tough and tricky properties of sequence data are their redundancies and discrepancies among resources of the same species or strain. They are brought in physically by different labs, depth of genome sequencing, draft forms of sequence assembly, gene prediction, genomic annotation, and so on, or genetically by mutant strains for microbial genomes, divergent or convergent evolutions, and so forth [17,18]. Since genes from these resources increase the difficulties in sequence processing, it will make a probe designer (usually a company's software for clients) difficult to process them and fulfil the criteria for oligo selection, which results in a considerably decreased probe productivity.

This study will address these issues as well as the problems with sequence similarity and homology that affect the quantity and quality of oligos. We introduced the concept of open reading frame (ORF) sequence segmentation from which quality oligos are to be selected, and we presented our algorithms and data processing strategies for gene sequence preparation. With increasingly massive sequence data, they need to be professionally handled and specially manipulated before submission to a probe designer that does not cope with these issues and pitfalls. It is also necessary to enable and facilitate a probe designer to select the most probes based on the major criteria by providing it with quality source sequences. Finally, we propose a probe database curation to keep track of experimentally workable probes in microbial communities.

2. Methods and Algorithms

The pan-genome sequence data used in this study consist of the finished core-genome genes from *Clostridium thermocellum* ATCC 27405 (DSM 1237; 1 major contig-based), strain-specific genes from the permanent draft genome *C. thermocellum* JW20 (DSM 4150; 21 major contigs-based) and mutant-specific genes from the draft genome *C. thermocellum* DSM 2360 (107 major contigs-based). Non-pangenome sequence data from a single genome *Thermobifida fusca YX* without mutant strains were used for a demonstration of results of a single-genome data processing. The numbers of oligos to be able to generate from these sequence data were shown with the probe designer eArray from Agilent. ORF analysis, cross-hybridization check, and pre-processing of all shared and strain-specific genes were carried out using our Perl-based pipeline ORF-Purger 2.0, which is available on request (rli8@hotmail.com).

2.1. Sequence data acquisition, collection and classification

Acquisition and collection of genome-scale sequences are made mainly from well-established and well-curated databases and Web pages from NCBI, JGI, etc. as far as microbial genomics is concerned; they also can be from user's own databases, ORF finding or gene predicting results, and whole shot-gun genome or individual contig annotations. These sequence data are of different types and are prepared by classifying them into three major groups: 1) The sequences from GenBank. These data have the first rank and are treated as a standard and termed "base data". Their data files with the name extension .ffn (fasta feature nucleotides) or .gbk (GenBank flat file or a Web page) contain all gene sequences and features annotated and extracted from a genome. Their gene features data given in the header line of a FASTA file take precedence to be kept over other sources of data. Among the gene features, especially, the gene symbol and locus (ORF coordinates) are pivotal items (right like primary keys in relational databases) to provide connections of one gene dataset with another. Since there are possible six reading frames along double DNA strands (three at positive strand and three at negative strand), there are possibly discrepant ORF coordinates from data files based on annotation of multiple contigs. They are essentially referring to the putative loci of the same gene found or predicted in different ways. For microbial genomes, there can be more reading frames along closed circular DNA strands and more ORF coordinates thereof. Only GenBank-based (also one major contig-based) ORF coordinates (gene locus) are kept and used as a criterion for later integration of gene symbols across data files. Likewise, there are variants of a gene symbol and only one from GenBank-based data file is kept as a benchmark. By convention, a gene symbol takes the form of a few initial letters of organism name, an underscore, and four digits. The base data with all these norms will be input first for processing. 2) The well-established or published sequences from those well-curated databases such as JGI-IMG. The data of this rank will be input after the base data but before the other lower rank data. 3) The draft-form sequences from draft genome assembly and/or annotations. The data of this rank will be input at last such that any unsuitable sequences will be eliminated from gene entries backwards. Inclusion of draft genomes is just for the purposes of collecting all potential genes available that would yield a maximum number of oligos in terms of probe integrity. Thereby, all the sequences for a pan-genome obtained from as many resources as possible constitute a "gene pool" that is as inclusive as possible for an integral collection of genes of interest. The experimental gene pool in this study are 18,810 ORFs (genes) input from a collection of 132 sequence files acquired from NCBI and JGI for the same species *C. thermocellum* upon which sequence data analysis and pre-processing are based.

2.2. Sequence data analysis, cleaning, purgation and integration

After sequence data are properly classified, they are analyzed by a global alignment (ORF-long base-to-base comparisons) and determined if they are representing the same gene with identical base composition. Once an identical ORF sequence is found (a gene match) by pairwise ORF alignment, the second ORF is dropped by not being included in the gene pool. However, the gene features data, if any, that come with the dropped ORF will be used to merge with that in the first ORF if it lacks in any data of corresponding gene features. The first ORF is termed "base ORF" as it'll be always kept after each paired comparison. For example, if the base ORF has no feature information of protein description (a gene product name) and the dropped ORF contains that information, it will be used to make up that deficiency of data in the base ORF. This way makes the base compositions of genes molecularly analyzed and genetically identified and the gene feature information biologically integrated. Any duplicate ORFs are properly filtered out such that the gene pool is cleared of any data redundancy no matter how many sequences are input and where they are collected [22]. And what is more, the missing data in the base ORF have been supplemented with otherwise available data. This improves sequence data integrity without losing any useful existing information based on a genetically secured ORF-to-ORF sequence correspondence (i.e., by the mutual nt base identity). After this preliminary processing, all raw data have been purged without redundant ORF (gene) entries (only one copy of each distinct gene in the gene pool), and cleaned without hidden spacings between bases and with all uppercased base letters. In addition, the base ORF (gene) entries have been integrated with complementary information from those removed ORF sequences.

2.3. Validation of ORF sequences

Generally, the gene sequences obtained from resources are those that have been output from ORF/gene finder such as GeneMark, Prodigal, and Glimmer or annotation systems like JGI-IMG [16,23]. By definition, each of such gene sequences should begin with a start codon, then ORF body, and end with a stop codon. A start codon is a triplet of three bases functioning as an initiator to start the protein translation process. Unlike in eukaryotes where ATG is the only initiator codon, there are alternative codons ending with the motif "TG" that can act as initiators in prokaryotes, resulting in a bunch of closely related sequences. This is one of factors that contribute to a diversity of putative genes and to the sheer number of gene families. For microbial genomes, a start codon could be ATG, GTG, TTG, CTG or, in rare cases, ATT; an ORF body is defined as the segment between start codon and stop codon, exclusive; and a stop codon is a triplet of thre bases (i.e., TAA, TAG, and TGA) functioning as a terminator to signal the end of protein synthesis. The full-length ORF consisting of these two codons and their intervening sequence (ORF body) constitutes an in-frame DNA sequence that begins reading a multiple of 3 bases until a stop codon is encountered. Some ORF definitions do not contain a stop codon in a given reading frame, but for the convenience of our study, a stop codon is treated as if it was a component of an ORF sequence throughout. With these definitions and conventions, inspecting the validity of an ORF sequence is performed by checking if it has a valid start codon of ATG, GTG, TTG, CTG, or ATT, a valid stop codon of TAA, TAG, or TGA, and a valid ORF length that should be greater than its oligo length by at least six bases (start + stop triplets). Any invalid ORF sequences found against these criteria are taken off from the gene pool.

2.4. Segmentation of ORF sequences

This is a very important operation we need to carry out in that quality oligos will be constructed from a redefined in-frame DNA segment. With the study subjects of microbial genomes, they posess a principal advantage, over eukaryotes, of the RNA splicing-free process as ORFs are transcribed. This makes alignment of an ORF with its transcript collinear without interference from uncertain gaps (introns). So the base composition of an ORF is in synteny with that of mRNA and an ORF sequence is synonymous with its protein sequence simply via direct translation. In this principle, any oligo defined straight from an ORF sequence will be of genetic significance to its target with which it is able to hybridize. By definition, a broad-sense gene is an ORF plus its upstream untranslated region (5'-UTR) and downstream untranslated region (3'-UTR). A narrow-sense gene is its ORF per se. Now totally for the purposes of constructing quality oligos from microbial genomes, an ORF itself should be subdivided into three segments: start codon, ORF body, and stop codon. Since stop codons do not code for an amino acid and hence are 'nonsense' to be included in an ORF sequence from which oligo is generated. Because all start codons are translated ultimately as one amino acid 'methionine' or, in bacteria, as “formylmethionine”, they are regarded as the “same thing” without contribution to differential gene products. That is, the “homogeneity” of start codons does not contribute to a variability of ORFs. Although a start codon has 'sense' to be included in an ORF sequence, its triplet is only limited to a few base compositions, namely ATG, GTG, TTG, CTG, or ATT among which ATG is most common, GTG and TTG are less frequent, and CTG and ATT are rare. Because of the restricted base compositions of start codons, they are not diverse enough to help generate gene-specific oligos if left in ORF. What is worse, chances are start codons in oligos may all be identical (like ATG) as long as they are included in probe selection (Table 4). It was known that the starting bases of an oligo probe take a crucial role in initiating a probe-target hybridization and determining the ensuing formation of hybrid duplexes [23,24]. Since an oligo is essentially just like a primer and “priming” a DNA hybridization kicks off at its starting bases or the 5’ end, all identical start codons (like ATG) in oligos can initiate free base pairing with any targets having complementary starting bases or the 3’ end. This reduces the specificity of oligo-target binding and is one of sources of potential cross-hybridizations. From the perspectives above, it is safe not to include these initiator codons in an oligo for a DNA microarray experiment.

There are two serious consequences of including start codons in an oligo sequence: 1) The same initiator codons contained in oligos are prone to non-target-specific hybridization initiation and 2) the presence of multiple or many identical start codons (like ATG) among oligos reduces the dissimilarity across oligos. This is undesirable if too many oligos have matchable or sharable base compositions [1]. Because of the nature of oligos that may contain identical start codons, they are not differential enough to ensure probe specificity and they may even get cross-hybridized simply from initiation, which makes start codons excluded from ORF sequences accordingly. For this reason, only genetically significant and differential segment of ORF body is considered to be a proper stretch of bases from which quality gene-specific oligos can be generated. Such a stretch of bases or in-frame DNA sequence without start and stop codons is the conceptual and logical segmentation of an ORF sequence and is still termed as ORFs (genes) hereinafter for short and convenience.

2.5. Purgation of identical gene sequences

After segmentation of ORFs, the shortened sequences (ORF body) relative to originally full-length ORFs are exposed to being possibly identical across ORFs when start and stop codons are stripped off. These shorter sequences are subject to a global alignment (ORF body-wide base comparison) again to detect ORF (gene) identity, and are purged once identical ones are found. Because only ORF bodies are regarded as distinct in-frame DNA sequences that contribute essentially to differential gene products, they are real unique segments of genes. That is, only one copy of such segment for each gene is kept after this purgation. The ORF body is treated as a unit of distinct stretches of bases and will undergo subsequent processing and be used as input sequence data for a probe designer. Again, the base ORF entries are updated by integrating with the complementary and meaningful information, if any, from those removed ORFs as compared to them.

2.6. Removal of overlapping gene sequences

Next we need to clear the gene pool of any overlapping ORFs in that any oligo selected from one overlapping ORF sequence will definitely belong to another ORF sequence with which it overlaps. An overlapping ORF is just like a subsequence of another longer ORF sequence and looks like something as follows, as shown by the bold and underlined fragments of DNA.

Simulated data:

ORF 1 (longer gene)
AGTTACAAATCGAATTTATCCAACT**GCAGCAAAAGTATTATCAACTCTATTA**ATCTTC
ATTTGG
ORF 2 (shorter gene) ==> GCAGCAAAAGTATTATCAACTCTATTA

Real gene data:

ORF 2360_5342 (longer gene)
GATGTACTCGGCCTTAAAGGAAAGAAAACAAAGGATGATGCCGAAAAATTTGTTTT
ACAATATATTAAAAATTCAGGCTTGTATATGGTATTATATACTTGTCAG
ORF JW20_3894 (shorter gene)
GATGTACTCGGCCTTAAAGGAAAGAAAACAAAGGATGATGCCGAAAAATTTGTTTTACA
ATATATTAAAAATTCAGGCTTGTATATGGTATTATATACTTGT

The existence of multiple reading frames, different algorithms used in gene prediction, and multiple contig annotations (especially for closed circular DNA species) may produce overlapping ORFs. These overlapping ORFs may not be detectable but appear after start and stop codons are trimmed, which is right another functionality of ORF segmentation that helps discover them. Overlapping ORFs are unwanted sequences that make oligo selection impossible for the shorter one and difficult for the longer one or have much less room left for oligo search, thus removing them improves the quantity and quality of oligos to be designed. Only the shorter one of two mutually overlapping ORFs is removed as it can be completely represented by the longer one. Take the above overlapping ORFs of ORF 2 and JW20_3894 for instance, they have no oligo selectable because all possible candidate oligo sequences extracted from them are falling within the regions of ORF 1 and 2360_5342, respectively. Thus, ORF 2 and JW20_3894 have no oligo to generate and need to be removed from the gene pool in order for ORF 1 and 2360_5342 to have one. Optionally, information about these dropped ORFs will be saved in a file and later used to make sure that those oligos constructed from ORFs like ORF 1 and 2360_5342 won't overlap with (fall within) the dropped ORF sequences like ORF 2 and JW20_3894.

2.7. Elimination of similar gene sequences of equal length

A similarity of gene sequences is abiotically due to a larger collection or superset of sequences from various resources that have pan-genomic data and/or different approaches for data acquisition and manipulation. It also can be caused phylogenetically by sequence homology, which is addressed next. For an individual data file or a single source (or type) of sequences for a particular organism/species without multiple strains, or with one approach to acquiring data, this is not a big issue. For such data, this step of purgation may not make much difference in improving the quantity and quality of probes to be designed. As genomic information keeps expanding, sequence data from a core genome combined with multiple related strains, mutants or isolates are available for a joint or comparative study. This pan-genomic data contain a high degree of sequence similarity shared by member strains apart from gene redundancies [18,22]. For such data, this step of purgation becomes significant in sequence processing, as similar sequences would make oligo selection difficult for a probe designer (Table 5). The idea behind this procedure is that presence of sequences of the exactly same length is a very rare event and that there is a little probability of a single base being mutable but little probability of two contiguous bases or a triplet of three bases (by entire codon) being mutable. The base mutability here can refer to transition or transversion or to abiotic alterability due to possible errors and/or discrepancies brought by human and programmed work. Therefore, any gene sequences having the same length and a small percentage of in-frame mismatched bases (one per codon) for whatever reasons could be considered the same gene. Their redundants are purged accordingly for a maximum number of oligos to be able to generate. For a large collection of datasets, gene sequences with highly mutual similarity look like something as follows (the mismatch is bold and underscored):

Real gene data (n = 243, m = 1, pct = 0.00411522633744856)
ORF 2360_5754
GCGGTAGATTACAATGCGATATCAAAAATATATGATAAAGTAAGGTCGGAAAACAAAA
ORF 642890680
CCGGTAGATTACAATGCGATATCAAAAATATATGATAAAGTAAGGTCGGAAAACAAAA

where only the first bases (G and C) of two ORF sequences mismatch and the rest of bases match; the length of either ORF = 243, the number of mismatches = 1, and the dissimilarity percentage = 0.004.

The ORF 642890680 shown has very limited room of searching for candidate oligos, which makes probe design difficult. Since both ORFs are of the same length, chances are they are representing the same gene and are derived from the originally common ancestor. They diverged from each other just because of at least one-time transition and/or transversion mutations between individual bases in the evolutionary process or simply due to errors and/or discrepancies introduced throughout the processes of genome sequencing, sequence assembly, gene prediction and contig annotation. Especially, these discrepancies could be aggravated when strain-specific genes are merged in a pangenome. Therefore, these similar ORFs are considered multiple copies of one gene without substantial differentiations between them. It would exacerbate the difficulties in probe selection if such a redundant ORF entry remains in the gene pool. Thus eliminating it (e.g., ORF 642890680) improves the quantity and quality of probes to be designed. The elimination percentage starts from 0.001 dissimilarity that is calculated by the number of in-frame mismatched bases (one per codon) of pairwise ORFs divided by their length. Practically, the stringency of this criterion can be relaxed by counting any mismatched bases as long as the processing result allows a probe designer to be able to select a maximum number of oligos.

2.8. Elimination of similar gene sequences of unequal length

A homolog for a gene within the same organism in datasets could occur either biotically or abiotically. Biotic homologs are derived phylogenetically from historical frameshift mutations, whereas abiotic homologs are probably the artifacts from different ways for data acquisition, especially when highly sharable genes are merged over strains. With sequence homology, it is hard for a probe designer to construct the most probes from closely related gene families while maintaining their higher specificity. Nevertheless, the overwhelming criterion for a better probe design is to achieve a probe integrity, whether proceeding to this step of purgation or not depends on if a maximum number of oligos can be produced by a probe designer. Most of the time, this homolog elimination may not be necessary unless the following scenario develops: The above similarity-based purgation does not improve a lot the productivity of oligos while the number of oligos has yet to be maximized.

Homolog elimination will not be based on the intervening homology or intergenic consensus that would be responsible for less clusterings of homologous sequences from each of which candidate probes are selected. Instead, use pairwise gene sequences of unequal length for homology analysis based on shorter-ORF-wide alignment. Their comparison begins from both starts through the end of the shorter ORF without alignment towards the end of the longer ORF. As described in the similarity case, different ORFs with equal lengths are most likely to refer to the same gene since the probability of equal-length genes is very small. However, it is also possible for unequal-length genes to refer to the same gene due to at least one-time base addition and deletion mutations in the lengthy evolutionary process and again errors or discrepancies introduced from human and programmed work. This treats each pair of the shorter and longer ORFs as the essentially same gene except the divergent lengths caused by these factors. That is, these homologous ORFs also can be deemed multiple copies of one gene without substantial differentiations between them if their homology is very strong by the small percent dissimilarity. Actually, all homologs will result in a reduced number of oligos to select without respect to levels of homology. Therefore, they are eliminable by a certain 257 pseudo-genes that had no valid start and/or in-frame stop codons. After that, six sequences with the in-frame DNA fragments shorter than the input sequence lengths were found, in which case the percentage of penalty until a maximum number of oligos can be generated from a probe designer. The elimination percentage starts from 0.001 dissimilarity that is calculated by the number of mismatched bases of pairwise ORFs with the length of the shorter ORF as a denominator.

2.9. Oligo probe check for sequence overlaps

An oligo (oligonucleotide) is a DNA/RNA fragment of 20–80 bases (or a 20–80-mer for a short nucleic acid polymer). For oligo probe-based DNA microarray designs, an oligo sequence is selected from a consecutive stretch of bases within an ORF sequence. If an ORF sequence has at least one selectable oligo sequence that does not overlap all other ORF sequences in the gene pool, it is considered an ORF sequence with an oligo available. If an ORF sequence has all oligo sequences selected, each of which overlaps all other ORF sequences in the gene pool, it is considered an ORF sequence without an oligo available. Assuming an oligo of 21-mer required to select, an illustration of the gene pool composed of five ORFs with or without an oligo is given below, as shown by the bold and underlined fragments of DNA:

ORF1: **ACAGGTTTGCGTCGGTTTACAGTC**
ORF2: TCTTATGTTA**GGTTTGCGTCGGTTTACAGTC**AGTTTAAGGATATA
ORF3: GCAAAGCT**CAGGTTTGCGTCGGTTTACAG**AACAAAC
ORF4: AAC**AGGTTTGCGTCGGTTTACAGT**ATGGACGAGCTTCGCAAGTTAGAG
ORF5: CGAAATGTGACTAAGG**ACAGGTTTGCGTCGGTTTACA**AGCTGATAGACACT

Oligo 1: ACAGGTTTGCGTCGGTTTACA
Oligo 2: CAGGTTTGCGTCGGTTTACAG
Oligo 3: AGGTTTGCGTCGGTTTACAGT
Oligo 4: GGTTTGCGTCGGTTTACAGTC

where a total of four possible oligos (Oligos 1, 2, 3, and 4) can be selected from ORF 1 but each of them overlaps all other four ORF sequences in the gene pool. Then ORF 1 is an ORF without oligo available or constructible. ORFs 2, 3, 4, and 5 have all of their oligos selected, each of which may overlap one or some of other four ORF sequences, or overlap none, thus they are ORFs with oligos available or constructible. Any of the ORFs that has no oligo constructible will be excluded from the gene pool, as inclusion of it in the input sequence data for a probe designer will do nothing but make the program take it as a possible candidate sequence. This not only takes up the time but exacerbates the difficulties in probe selection when that oligo-free ORF is taken into account as a whole and for a trade-off. Therefore, to avoid a poor candidate probe search, the gene pool should be purged, as a last step, of any oligo-free ORFs, which makes all the genes that are subject to all pre-processing the proper and sound source of input sequence data for a probe designer.

3. Results

The sequence data subjected to the required or basic pre-processing were exhibited in Table 1 where 12,210 duplicate sequences were purged by global alignment from the gene pool of 18,810 ORF (gene) entries that ended up with 5,562 quality genes. Validation of ORF sequences detected 257 pseudo-genes that had no valid start and/or in-frame stop codons. After that, six sequences with the in-frame DNA fragments shorter than the input sequence lengths were found, in which case the extra bases beyond the reading frames could be caused by errors or algorithms in gene predictions or genome annotations (Table 2). Five pairs of aligned sequences with the same ORF bodies except start or stop codon exemplified that genes could differ from one another just by the start or stop codon (Tables 3).They are also examples in the extreme case of producing closest related gene

Table 1. A stepwise sequence purgation for required or basic pre-processing of sequence data from a pan-genome of the organism *Clostridium thermocellum* with a huge collection of 132 sequence files.

Steps of Purging Sequences	Results of Sequences	Number of Sequences
Input ORFs from the gene pool	Total ORF sequences	18810
Purgation of redundant ORF sequences	Duplicate sequences	12210
	Distinct sequences	6600
Validation of ORF (gene) sequences	Genes shorter than 60 bp*	11
	Pseudo-genes	257
	Validated genes	6332
Purgation of identical gene sequences after undergoing the ORF segmentation	Identical genes	11
	Distinct genes	6321
Removal of overlapping gene sequences	Overlapping genes	446
	Non-overlap genes	5875
Elimination of similar gene sequences of equal length	Not used for basic processing	------
Elimination of similar gene sequences of unequal length	Not used for basic processing	------
Oligo probe check for sequence overlaps	Genes w/o oligos available #	313
	Genes with oligos available	5562
End of sequence pre-processing	Output unique/quality genes	5562

* The oligo probe size for 60 base pairs or 60 mer. It is subject to change and used for gene size check.

# The number of genes without oligos available will become smaller when the step of Elimination of similar gene sequences of equal length (and/or unequal length) is taken.

families of two members each. Together with the other extent of similarity within ORF body, this accounts partially for many closely related gene families in a very large genome or pan-genome.That makes probe design difficult. Hence 11 pairs of genes (6 from Table 2 and 5 from Table 3) were treated as two identical sequences each and the second one was purged. Some of the resultant oligos generated by the probe designer eArray using sequence data without segmentation processing of ORFs (i.e., both start and stop codons were not stripped off) were illustrated in Tables 4. It showed the worse consequences of the multiple oligos starting with ATG, which could initiate cross-hybridization and reduce the dissimilarity across oligos.

Table 2. Six pairs of aligned genes with the same in-frame DNA but different input sequence lengths.

Gene_ID	Genes with the same reading frames
JW20_4577	**ATG**GATAAGTTTATTAAACAATTAGATCCAAACTTA//AACACCGGTTGT**TGA**
640278321	**ATG**GATAAGTTTATTAAACAATTAGATCCAAACTTA//AACACCGGTTGT**TGA**CAAA//AAC**TAA**
JW20_4063	**ATG**ATAATCAAATGTGAAAAAATTGTAAAGACTTGT//CCAGGGCTCTTA**TAA**
640278319	**ATG**ATAATCAAATGTGAAAAAATTGTAAAGACTTGT//CCAGGGCTCTTA**TAA**GCAA//GCA**ATA**
2360_5493	**ATG**CTGCAAATTGCGCTCTGCGATGACAATACAAAC//AAGCAGCGGATT**TGA**
640278317	**ATG**CTGCAAATTGCGCTCTGCGATGACAATACAAAC//AAGCAGCGGATT**TGA**ATTA//GAG**TAA**
2360_5748	**GTG**TCAACATTATCAAAAGAGCAAGTGAAAGAAATA//GATGAAATTCAC**TAG**
640278310	**GTG**TCAACATTATCAAAAGAGCAAGTGAAAGAAATA//GATGAAATTCAC**TAG**ACAA//AGA**TAA**
2360_5770	**ATG**TCAACATTATCAAAAGAACAAGTAAAAGAAATA//AGAATAGCACCT**TGA**
640278301	**ATG**TCAACATTATCAAAAGAACAAGTAAAAGAAATA//AGAATAGCACCT**TGA**GATT//ATT**TAA**
2360_5274	**ATG**TTCTTTGATAAAGGAGATTACTCAGAGCACAAC//CCTGCAAAACTG**TAG**
640278294	**ATG**TTCTTTGATAAAGGAGATTACTCAGAGCACAAC//CCTGCAAAACTG**TAG**CAGC//AAA**GGT**

The significance of sequence data pre-processing using the elimination of similar genes of equal length was shown in Table 5, which further purged 1,620 genes from the gene pool and ended up with 4,240 optimal genes. The performance of sequence data pre-processing for improving the quantity and quality of oligos was demonstrated in Table 6 where the oligos with potential cross-hybridization were detected from the output probes generated by eArray and the unique oligos were evaluated. For sequence data without pre-processing, only 225 unique oligos, the smallest probe quantity, were available in the microarray experiment for *C. thermocellum* and 46 for *T. fusca YX*, making them useless whatever their data types/sizes would be. For sequence data with basic pre-processing, there were 2,744 unique oligos, the drastically elevated probe quantity, available in the microarray experiment for *C. thermocellum*, making it marginally applicable, and 3,111 for *T.*

fusca YX, making it applicable. Probe productivity from a single-genome was superior to that of pan-genome in that the former had less sequence redundancies and similarities. For sequence data with optimal pre-processing, there were 3,615 unique oligos, the further elevated probe quantity, available in the microarray experiment for *C. thermocellum*, now making it applicable, and 3,111 for *T. fusca YX*, making no difference. Namely, elimination of similar genes of equal length had less or no effect on the probe productivity from a single genome if it was sufficiently dissimilar over genes. For the best probe integrity, those 3,615 oligos were well reflected to be closer to the gene count 3305 (NCBI) or 3335 (JGI) from one major contig-based core-genome annotation, plus a reasonably inceased number of other strain-specific genes. Those 3,111 oligos were well reflected to be closer to the gene count 3184 (NCBI) or 3195 (JGI) from one major contig-based single-genome annotation. Excluding RNA-coding, regulator, or house-keeping genes, as suggested by the CDS count 3110 (NCBI) or 3117 (JGI), the probe integrity was warranted and just about satisfied.

Table 3. Five pairs of aligned sequences with the same base compositions except start or stop codon.

Gene_ID	Gene sequences with identical ORF bodies
27405_1409	**TTG**AAAACGTGTATTGCTTGTGGAATGCCTATGAAGGATATTTCAGACTTT//TATTTAAT**TAA**
2360_5735	**ATG**AAAACGTGTATTGCTTGTGGAATGCCTATGAAGGATATTTCAGACTTT//TATTTAAT**TAA**
JW20_4146	**ATG**GCAGGGGATCATTTATATGCTCCGTTTGTGGAAAGTGTAAAAAAAGTT//ACTGACAAG**TAA**
2360_5047	**ATG**GCAGGGGATCATTTATATGCTCCGTTTGTGGAAAGTGTAAAAAAAGTT//ACTGACAAG**TAG**
27405_2041	**ATG**AGCTTTTATGCCGCACCGATTGCAAGGCTTATAGAGGAGTTTGAGAAG//CGCGAGATT**TAG**
JW20_4452	**ATG**AGCTTTTATGCCGCACCGATTGCAAGGCTTATAGAGGAGTTTGAGAAG//CGCGAGATT**TAA**
27405_2221	**ATG**AGCGCGAAAATCCTTGTTGTTGATGACGAGAAAAATATAGTTGACATT//AAATTAAGT**TAA**
JW20_4195	**GTG**AGCGCGAAAATCCTTGTTGTTGATGACGAGAAAAATATAGTTGACATT//AAATTAAGT**TAA**
27405_2542	**TTG**TGGGTATCTGTTAGCAATCAGGCATATGTTTTTTTAAATTGTGTTCTC//AAAAAAATA**TAA**
JW20_4049	**TTG**TGGGTATCTGTTAGCAATCAGGCATATGTTTTTTTAAATTGTGTTCTC//AAAAAAATA**TAG**

Table 4. Oligo probes generated from ORFs of full length (start codon + ORF body + stop codon) by the probe designer eArray with start codons being selected from the first base location (BL = 1) of DNA strands from the pan-genomic sequence data of the organism *Clostridium thermocellum*.

Gene_ID	BL	Oligo probe sequence (60mer)
640106511	1	**ATG**AGTGTTTTAAAGGTTTCAGCAAAATCCAATCCAAATTCCATAGCAGGTGCTTTGGCG
640107014	1	**ATG**GCAGATACTTTGGGCAGGGACTTTCTTAGAGCATTGTTCAAACTAAAAAGCTTGCTT
640107063	1	**ATG**AGTATCAGCAAGTTTAACGCAGAAGGATATTACGACCCCACACCTTATGAAGCACTG
642888068	1	**ATG**GCAGATACTTTGGGCAGGGACTTTCTTAGAGCATTGTTCAAACTAAAAAGCTTGCTT
642890610	1	**ATG**AGTAAGAAGAAGTACATTGTCCGCTGTCCTCACTGCAATCACAGAGTATTTGATGCC
2360_1869	1	**ATG**TTCACATGGGATACATCGGTATATCCGTGGGTTCGTTTTTATCGTACCTATGAGGAA
642888280	1	**GTG**AAGCAACAATTCAAGTTGAAGCCTATGAGACACCAATACCGAAAAGTAATCAGGTCG
642888463	1	**TTG**AGCAATATGGTGTACGACATTCTTGAAAGCCAGATAAAGAACGAGATCGACAGAAGT

Table 5. A stepwise optimized sequence purgation following required or basic pre-processing of sequence data from a pan-genome of the organism *Clostridium thermocellum* with 132 sequence files.

Steps of Purging Sequences	Results of Sequences	Number of Sequences
Follow Removal-of-overlap step	Non-overlap genes	5875
Elimination of similar gene sequences of equal length*	Similar genes (purged)	1620
	Singleton genes (kept)	4255
Oligo probe check for sequence overlaps	Genes w/o oligos available	15
	Genes with oligos available	4240
End of sequence pre-processing	Output unique/optimal genes	4240

* The optimized sequence purgation was met by similarity elimination at percent dissimilarity = 0.01.

Table 6. A demonstration of performance in sequence data pre-processing that resulted in improvement on the quantity and quality of oligos generated by the probe designer eArray using the pan-genome data of the organism *Clostridium thermocellum* in contrast to a single genome of *Thermobifida fusca YX.*

Sequence data and the resulting probe quantities without pre-processing

Sequence Data Types	Input raw sequences	Output oligos	X-hyb oligos*	Unique oligos
Pan-genome ORFs from *C. thermocellum*	18810	6580	6355	225
Single-genome ORFs from *T. fusca YX*	6305	3172	3126	46

Sequence data and the resulting changes in probe quantities with basic pre-processing

Sequence Data Types	Input raw sequences	Quality genes	Output oligos	X-hyb oligos*	Unique oligos
Pan-genome ORFs from *C. thermocellum*	18810	5562	5558	2814	2744
Single-genome ORFs from *T. fusca YX*	6305	3115	3115	4	3111

Sequence data and the resulting changes in probe quantities with optimal pre-processing

Sequence Data Types	Input raw sequences	Optimal genes	Output oligos	X-hyb oligos*	Unique oligos
Pan-genome ORFs from *C. thermocellum*	18810	4240	4237	622	3615
Single-genome ORFs from *T. fusca YX*	6305	3115	3115	4	3111

* Oligo probes that could potentially cross-hybridize to non-target sequences present in the hybridization reaction.

4. Discussion and Conclusion

In light of the best probe integrity as we introduced in this study, optimization of a probe design is not only focused on the improved specificity of oligos, one of the major criteria for quality probe selection, but also on the improved quantity of oligos. The specificity could be made better by choosing a longer probe such as 60 mer through 80 mer, which is not necessarily compromising the sensitivity to hybridization, as long oligos have been shown to be more sensitive than shorter ones [3,10]. The main questions with longer oligos are that they have the potential of self-complementarity (self-hybridization) or, what is worse, the formation of stable secondary structures. As such, a suitable oligo length could be adjusted to make probes as much gene- or sequence-specific as could, which is generally able to be fulfilled [5,10,25]. Our experiments also indicated that the overall specificities of all oligos were satisfied without respect to levels of homology or similarity of partial base compositions. On the other hand, there is a trade-off between the specificity and quantity of oligos: the higher the former is, the less the latter would be, or vice versa. Because of the necessity of full availability of all possible probe-target bindings, the specificity and quantity of oligos are of equivalent importance as to minimize cross-hybridization. When a complete set of oligos are available, each oligo has a full complementary relation and interaction with its target, which preempts a direct, strong, and stable probe-target binding or DNA duplex with the highest affinity, thereby precluding probe-nontarget cross-hybridization. When an incomplete set of oligos are available only, a probe could bind to a nontarget if the affinity of the nontarget to bind to its own probe is absent because no probe can be found for this nontarget. The quantity of oligos could become problematic for very large microbial genomes (> 5,000 genes), especially when pan-genome sequence data of higher redundancies are used for probe design. They usually result in a dramatically smaller number of oligos relative to their sheer size if processed by a probe designer. Thus, the overwhelming criterion for a better probe design is to achieve a probe integrity, with the less stringency of other criteria for probe specificity and sensitivity under the given design constraint on thermodynamic uniformity at the hybridization temperature. From the philosophy of availability of full entity of probes and information integrity, presence of a good-quality but incomplete set of oligos is worse than presence of a moderate-quality but complete set of oligos. That is, the current specificity and sensitivity of longer probes are working but results are far from being informative due to a reduced number of probes. It is necessary for a full knowledge acquisition system through global gene expression analysis monitored by all possible hybridization signals. Particularly, it is also required for new gene and its function discovery, as suggested by a large number of hypothetical proteins in their entries of genes predicted [19]. Therefore, pre-processing of gene sequences using the segmentation and purgation of ORFs is necessary to prepare for molecularly valid and genetically significant data that allow optimization of probe design in terms of probe integrity. This integrity with an inclusive probe set will not only make experimental results globally informative but also help minimize cross-hybridization in the presence of a full set of probe-target correspondence [26].

We also propose a probe database curation to monitor, maintain, and manage all oligos that prove to be workable as probes in successful microarray experiments without redesigns in microbial communities [22]. This is due to the property and practical complexity of an interactive probe-target binding affinity and the performance of a target-specific hybridization probe may not be predicted or expressed as a high degree of sequence specificity. Their interplay might also be involved in thermodynamics other than free energy change [27,28] and a theoretical design may not receive the better results than its experimental approaches. For instance, a probe could be of high specificity but of low sensitivity or of other unknown poor affinity that still limits its availability for binding to its target [11,29]. Out of this thought, a heuristic solution to great demand for required yet robust probes is to keep track of experimentally workable or empirically functional probes and mark them as entries that do not need redesign in a new probe design. This way will facilitate a probe designer to generate the most probes in the challenging demand for probe integrity by skipping those entries without a tough trade-off consideration for them as a whole.

Because of the necessity of fully physical availability of all possible probes corresponding to their target sequences in a nucleic acid sample, the criteria for oligo probe selection in microbial genomes are set as (in the order of their relative importance): Probe integrity for target space (set), homogeneity in melting temperature (Tm), specificity to a single target, sensitivity to hybridization, self-complementarity, repeated region and distance from the 3' end [30]. Probe integrity is given a top priority in that having all possible probes present in hybridization reactions is better than any absence of them even at the sacrifice of Tm uniformity from the aforementioned philosophy. From the purely pragmatic standpoint of having a custom solution to probe design for very large microbial genomes or pan-genome sequence data, our future work will be done with a standalone program OliGo 1.0. It will design probes under the constraint that these criteria can be relaxed to select the most probes in addition to implementing the algorithms that are able to solve the problems presented in this study.

Conflict of Interests

All authors declare no conflicts of interest in this paper.

References

1. Kane MD, Jatkoe TA, Stumpf CR, et al. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays. *Nucleic Acids Res* 28: 4552–4557.
2. Rahmann S (2002) Rapid large-scale oligonucleotide selection for microarrays. *In Proc IEEE Comput Soc Bioinform Conf* 1: 54–63.
3. Russell R (2003) Designing microarray oligonucleotide probes. *Brief Bioinform* 4: 361–367.
4. Reymond N, Charle H, Duret L, et al. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics* 20: 271–273.
5. He Z, Wu L, Fields MW, et al. (2005) Use of microarrays with different probe sizes for monitoring gene expression. *Appl Environ Microbiol* 71: 5154–5162.
6. Li X, He Z, Zhou J (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res* 33: 6114–6123.
7. Li F and Stormo GD (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 17: 1067–1076.
8. Nielsen HB, Knudsen S (2002) Avoiding cross hybridization by choosing nonredundant targets on cDNA arrays. *Bioinformatics* 18: 321–322.
9. Krause A, Krautner M, Meier H (2003) Accurate method for fast design of diagnostic oligonucleotide probe sets for DNA microarrays. *IPDPS*: 1–9.
10. Letowski J, Brousseau R, Masson L (2004) Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. *J Microbiol Meth* 57: 269–278.
11. Nordberg EK (2005) YODA: selecting signature oligonucleotides. *Bioinformatics* 21: 1365-1370.
12. Jourdren L, Duclos A, Brion C, et al. (2010) Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments. *Nucleic Acids Res* 38: e117.

13. Rouillard JM, Herbert CJ, Zuker M (2002) OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* 18: 486–487.
14. Sung W, Lee W (2003) Fast and accurate probe selection algorithm for large genomes. *In Proc IEEE Comput Soc Bioinform Conf* 2: 65–74.
15. Hyyrö H, Juhola M, Vihinen M (2005) Genome-wide selection of unique and valid oligonucleotides. *Nucl Acids Res* 33: e115.
16. Markowitz VM, Chen IA, Palaniappan K, et al. (2010) The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* 38: D382–D390.
17. Oh S, Yoder-Himes DR, Tiedje J, et al. (2010) Evaluating the performance of oligonucleotide microarrays for bacterial strains with increasing genetic divergence from the reference strain. *Appl Environ Microbiol* 76: 2980–2988.
18. Hug LA, Salehi M, Nuin P, et al. (2011) Design and verification of a pangenome microarray oligonucleotide probe set for dehalococcoides spp. *Appl Environ Microbiol* 77: 5361–5369.
19. Markowitz VM, Mavromatis K, Ivanova NN, et al. (2009) IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25: 2271–2278.
20. Davidsen T, Beck E, Ganapathy A, et al. (2010) The comprehensive microbial resource. *Nucl Acids Res* 38: D340–D345.
21. Rimour S, Hill D, Militon C, et al. (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics* 21: 1094–1103.
22. Rouillard JM, Gulari E (2009) OligoArrayDb: pangenomic oligonucleotide microarray probe sets database. *Nucleic Acids Res* 37: D938–D941.
23. Hyatt D, Chen GL, LoCascio PF, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.
24. Wu C, Carta R, Zhang L (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res* 33: e84.
25. Hu G, Llinás M, Li J, et al. (2007) Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. *BMC Bioinformatics* 8: 350.
26. Flikka K, Yadetie F, Laegreid A (2004) XHM: A system for detection of potential cross hybridizations in DNA microarrays. *BMC Bioinformatics* 5: 117.
27. SantaLucia J J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95: 1460–1465.
28. Binder H, Preibisch S, Kirsten T (2005) Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir* 21: 9287–9302.
29. Binder H, Kirsten T, Loeffler Met, et al. (2004) Sensitivity of microarray oligonucleotide probes: variability and effect of base composition. *J Phys Chem B* 108: 18003–18014.
30. Liebich J, Schadt CW, Chong SC, et al. (2006) Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. *Appl Environ Microbiol* 72: 1688–1691.