CHAPTER 15

GENOME-WIDE CDNA OLIGO PROBE DESIGN AND ITS APPLICATIONS IN *SCHIZOSACCHAROMYCES POMBE*

Kui Lin
Beijing Normal University
linkui@bnu.edu.cn

Jianhua Liu
Genome Institute of Singapore
liujh@gis.a-star.edu.sg

Lance Miller
Genome Institute of Singapore
millerl@gis.a-star.edu.sg

Limsoon Wong
Institute for Infocomm Research
limsoon@i2r.a-star.edu.sg

Microarrays are glass surfaces bearing arrays of DNA fragments—also known as "probes"—at discrete addresses. These DNA fragments on the microarray are hybridized to a complex sample of fluorescently labeled DNA or RNA in solution. After a washing and staining process, the addresses at which hybridization has taken place can be determined and the expression level of the corresponding genes derived. Today, a single microarray can contain several tens of thousands of DNA fragments. Thus, microarrays are a technology for simultaneously profiling the expression levels of tens of thousands of genes in a sample.[518, 698]

In this chapter, we present a method for selecting probes to profile genome-wide gene expression of a given genome. We demonstrate our method on the genome of *Schizosaccharomyces pombe*. *S. pombe* or fission yeast is a single-celled free living Ascomycete fungus with many of the features found in the cells of more complicated eukaryotes. *S. pombe* is the second yeast, after *S. cerevisiae*, whose genome has been completely sequenced. Due to the fact that *S. cerevisiae* has undergone genome duplication and gene lineage loss and diver-

gence, *S. pombe* can be a better model organism for the study of gene expression, especially for those genes whose products are not present in *S. cerevisiae*.

ORGANIZATION.

Section 1. The biological background of *Schizosaccharomyces pombe* is summarized.

Section 2. The problem of designing oligo probes for genome-wide gene expression profiling by microarrays is formalized.

Section 3. An overview of our approach to solving this problem is then given, and its advantages briefly discussed. The approach has three modules. The first module extracts the coding regions from the given genome. the second module produces candidates oligos from the coding regions satisfying certain constraints such as G+C content, cross homology, *etc.* The third module selects from the amongst the candidates an optimal probe set satisfying some additional criteria such as minimizing distance to the 3' end of genes.

Section 4. The detailed implementation is then presented, such as aspects of data schema, object creation, criteria for probe production, probe production, and optimal probe set selection.

Section 5. The program is then run on the *S. pombe* genome to design a set of oligo probes for genome-wide gene expression profiling of *S. pombe*. The quality statistics of the resulting probe set is reported.

1. Biological Background

Schizosaccharomyces pombe was first isolated from an East African millet beer, called Pombe. It lives mainly as a haploid, divides by cell fission, and responds to nutrient starvation by mating to a partner with opposite mating type and forming four-spored asci. Its diploid cells can be maintained in the laboratory if zygotes are transferred to a rich medium. Like its distant relative, *S. cerevisiae*, *S. pombe* is amenable to genetic, biochemical, cellular, molecular, and functional genomic studies. *S. pombe* has served as an excellent model organism for the study of cell-cycle control, mitosis and meiosis, DNA repair and recombination, and checkpoint controls important for genome stability.

S. pombe, normally a haploid, spends most of its time in G2 and controls its cell cycle by regulating the G2-M phase transition. By contrast, *S. cerevisiae*, normally a diploid in the wild, has a long G1 phase and the major decision point for its cell-cycle entry occurs at the G1-S phase transition. Thus, both yeast species have provided important contributions to the discovery of the basic mechanisms of cell division.

Both *S. pombe* and *S. cerevisiae* have the same total DNA content, but *S. cerevisiae* divides its genome amongst 16 chromosomes, while *S. pombe* has just three.

That may explain why *S. cerevisiae* has simple structures of DNA-replication origins and centromeres, whereas *S. pombe* contains relatively complex organization of DNA-replication origins and centromeres, although not as complex as that found in higher eukaryotes.

S. pombe is the sixth eukaryotic genome to be sequenced,[898] following *Saccharomyces cerevisiae*,[292] *Caenorhabditis elegans*,[128] *Drosophila melanogaster*,[7] *Arabidopsis thaliana*,[31] and *Homo sapiens*.[467] The comparison of *S. pombe* and *S. cerevisiae* revealed that the genome of *S. cerevisiae* was more redundant and underwent lineage-specific gene loss. This makes *S. pombe* a more attractive model organism for functional genomic studies.

Some gene sequences are as equally diverged between the two yeasts as they are from their human homologs, probably reflecting a more rapid evolution within fungal lineages compared with those of the Metazoa. Due to the fact that *S. cerevisiae* has undergone genome duplication and gene lineage loss and divergence, *S. pombe* can be a better model organism for the study of gene expression, especially for those genes whose products are not present in *S. cerevisiae*.

2. Problem Formulation

The design of oligonucleotide probes for genome-wide microarray gene expression analysis can be formulated as follows. Assume that there are m protein-coding genes annotated in a completely sequenced genome, denoted here as $G = \{g_i : |g_i| \geq L,\ 1 \leq i \leq m\}$ where L is the minimum length of a gene considered and is set to 50 nucleotides. The goal of the design of oligonucleotide probes for genome-wide microarray gene expression analysis is to identify, for each gene g_i of length $|g_i|$ nucleotide bases, at least one sequence segment $g_i^{u,v}$ between positions u and v of g_i such that

- $1 \leq u, v \leq |g_i|$,
- $v - u + 1 = L$, and
- $g_i^{u,v}$ is specific to g_i only.

The specificity (*i.e.*, uniqueness) of one segment of DNA sequence is defined as the similarity between that segment and any portion of any g_j of the same length L, for $1 \leq j \leq m$ and $j \neq i$. The degree of specificity must be kept below some threshold specified *a priori* under some predefined similarity measure method. Typically, different oligo design softwares rely on different similarity measures. For instance, PRIMEGENS[911] uses the minimal length and expectation value of the local alignment algorithm BLASTN[23] to measure the similarity between two DNA sequences.

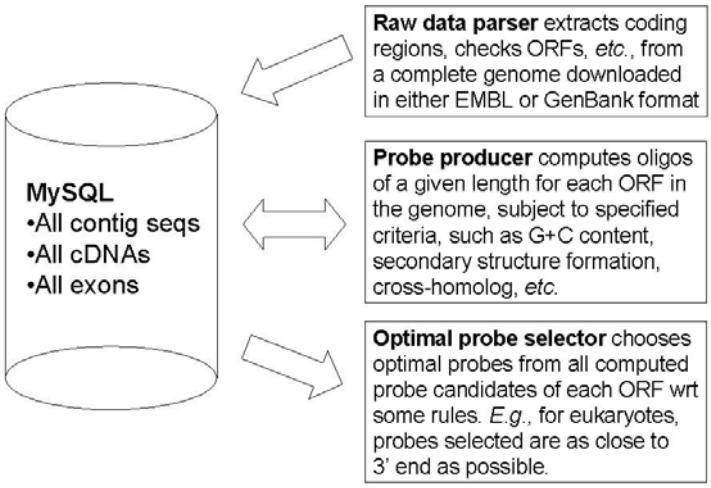

Fig. 1. The main modules of our oligo probe design program are the raw data parser module, the probe producer module, and the optimal probe selector module.

In our algorithm to be described shortly, the similarity measure used is called Hamming distance,[317] which is a measure of the specificity between two equal length DNA sequences. The Hamming distance $D_H(s_1, s_2)$ between two sequences s_1 and s_2 is defined as the number of mismatched nucleotides in the alignment of s_1 and s_2. For each probe candidate computed, our algorithm keeps the minimum Hamming distance that indicates how specific the probe is. The larger the minimum Hamming distance, the better the probe candidate is.

3. Algorithm Overview

We develop our oligo probes design program using the C programming language and using the MySQL relational database.[917] MySQL is used as a backend data warehouse for the storage of all raw genomic data, some intermediate data, and all final result. Figure 1 presents the main modules—raw data parser, probe producer, and optimal probe selector—of our program for the design of DNA oligo probes of *S. pombe*.

While our C-plus-MySQL framework increases IO operation time, we find that it also has many advantages. Firstly, we can integrate a large amount of related information that is important for computing a more optimal set of oligo probes for

the complete genome. For example, we can use exon information to avoid those probes whose locations span two contiguous exons in one ORF. This increases hybridization sensitivity in expression experiments.

Secondly, we are interested to parallelize the computation of oligo probe design for complete genome that may consist of more than several thousands of ORFs. We think it is easier to implement parallel computation on top of this C-plus-MySQL framework while maintaining consistency. For example, one way to accomplish this is to lock records or tables when we produce a new putative probe and insert it into specific table(s) for each individual process/thread running concurrently. See Figure 2.

Thirdly, this framework allows us to evaluate the quality of the selected set of probes. Having kept many intermediate results in the database, we can easily trace the hybridization result of each probe to the computed and logged information to see which factors are most related to the hybridization results, especially for those with "bad" hybridization signals. Are these "bad" results due to the "bad" selection of these probes? Or do they have a biological or experimental origin in for example, sample preparation, mRNA labeling, hybridization control, *etc.*? This analysis allows us to improve on the quality of our probe design in future experiments.

Lastly, as the probe selection computation is a time-consuming job in terms of CPU, it is advantageous if the program can be interrupted at any time, and be resumed at a later time if necessary. This functionality is very useful in situations such as when the server needs to be shutdown for system maintenance, or when the heavy load our of program is interfering with programs of other users, and so on. Under our C-plus-MySQL framework, the program can be easily interrupted and resumed according to the requirements of computational environment during the course of computing a microarray probe design.

4. Implementation Details

In this section, we present the details of our implementation, including aspects of data schema, object creation, criteria for probe production, probe production, and optimal probe set selection.

4.1. *Data Schema*

In our program, there are at least 5 core tables that must be created on the MySQL relational database before the probe computation can be carried out. We describe them in details as follows:

- `Contig` table stores all chromosomes, contigs, cosmids, and plasmids in the given complete genome.
- `CDS` table stores all ORF and CDS information, including location, strand, cDNA sequence, title, and number of probes computed so far, and so on.
- `Exon` table stores all exon information for each ORF.
- `Region` table holds specific information of each ORF and is used to help pick the candidate segment for the next round of computation with respect to pre-defined rules.
- `Probe` table stores all computed oligo probes. Each record contains useful information related to the probe, including G+C content, minimum Hamming distance, maximum number of contiguous nucleotide, location in the ORF, *etc*.

4.2. *Data Objects Creation*

A global array of `Contig` objects is designed to hold all contigs parsed from a given annotation file, which serves as the starting point of the design. The following segment is excerpted from the feature table of the annotated file of *Schizosaccharomyces pombe* chromosome I from EMBL database. It describes that a gene, sod2, consists of two exons which are located from nucleotide 62341 to 62464 and 62542 to 63824 in chromosome I. It encodes a protein called na(+)/h(+) antiporter in *S. pombe*.

```
FT  CDS         join(62341..62464,62542..63824)
FT               /gene="sod2"
FT               /note="SPAC977.10, len:468"
FT               /codon_start=1
FT               /label=sod2
FT               /product="na(+)/h(+) antiporter"
FT               /protein_id="NP_592782.1"
FT               /db_xref="GI:19113694"
FT               /db_xref="SWISS-PROT:P36606"
```

Note that, in our program, a contig represents an individual genomic sequence that is assembled in a given genome. It could be a chromosome, a contig, a cosmid, or a plasmid depending on the types of genomic sequences. Each `Contig` object in C is defined as follows:

```
struct Contig                   /* Contig object */
{
   char id[15];                 /* Identifier of the contig */
   unsigned long len;           /* DNA sequence length */
   unsigned int cdsNum;         /* number of CDS the contig has */
   char * title;                /* contig description */
   char * seq;                  /* DNA sequence */
   struct CDS * cds; /* pointer to CDS list */
};
```

Each `Contig` may contain one or many CDS/ORFs, which are organized as a single linked list in the `Contig` object, and each CDS/ORF represents a coding

region in the `Contig`. We define ORF and CDS objects in C as follows:

```
struct CDS                      /* CDS object */
{
   char id[21];                 /* Identifier of the CDS */
   char * title;                /* gene/ORF's name */
   int complement;              /* is it complement? 1=yes, 0-no */
   struct Exon * exons;         /* pointer to the exon list */
   unsigned long start, end, len;   /* ORF location */
   unsigned char * cDNA;        /* pointer to cDNA sequence */
   unsigned int probeNum;       /* number of probes computed so far */
   char isDuplicated;           /* Is it duplicated ORF? */
   unsigned long codons[62];    /* 61 codons and their sum */
   struct CDS * next;           /* pointer to the next CDS */
};
```

`Exon` objects are defined so that they can be used to select more robust or specific oligo probes for each ORF. For example, select a probe that do not span two contiguous exons of the ORF. We define `Exon` objects as follows.

```
struct Exon                     /* exon object */
{
   unsigned long start, end;    /* location */
   int phase, endPhase;         /* for splicing site */
   struct Exon * next;
};
```

A `Probe` object is defined to keep as much information pertaining to its computation as possible. Such information is important in selecting an optimal set of probes for each ORF. We define `Probe` objects as follows.

```
struct Probe                    /* oligo probe object */
{
   unsigned char * nt;          /* pointer to the probe size */
   int maxC;                    /* maximum contiguous number */
   int Tm;                      /* melting temperature */
   int GC;                      /* G+C content */
   int minHD;                   /* minimum Hamming distance */
   unsigned long start;         /* start position at the cDNA sequence */
};
```

When the input annotation file is parsed by the raw data parser module, an array of `Contig` objects is created in memory and all `CDS` objects—corresponding to ORFs in a contig—are also created as a single linked list that is linked to its corresponding `Contig` object. At the same time, all `Exon` objects belonging to each `CDS` are also created and organized as a single linked list which is attached to the `CDS` automatically. After having done some necessary checking work—such as start codon, stop codon, and so on—all types of objects are automatically imported into their specified MySQL tables whose schema are identical to the corresponding objects in C. Then the program terminates normally from the raw data parser module and the oligo probe production module can be launched.

4.3. *Criteria for Probe Production*

Before discussing probe production, we need to spend some time to characterize the criteria that the program uses in the oligo probe production module. Different strategies can be used in the computation of putative probes—*e.g.*, starting from the 3' end of each ORF first, or from some specific regions in the ORF at first. Regardless of the strategiy used, the selected segment needs to be checked against predefined constraints before it can be compared to other equal size segments from all the other ORFs.

Different programs typically use different constraints. However, the following constraints are usually taken into account in many probe and primer design programs:[409, 490, 518, 694]

- Each probe should have minimal secondary structure.
- Each probe should have no contiguous complementary stretches $>$ 15bp.
- Each probe should have minimal distance from the 3' end, taking into account poly-A prediction if possible.
- The combination of probes should be a maximal representation of alternative splice variants if possible.
- The combination of probes should avoid cross-hybridization.
- The combination of probes should have homogeneity in G+C content and melting temperature

It has been reported that, for good gene specificity, non-target cDNAs ought to be less than 75% in sequence similarity compared to the target region (50bp in size) to prevent significant cross-hybridization.[409] In our program, we define the following global variables to characterize the constraints on the oligo probes that we want:

```
MAX_Tm = 65;                    /* 1. max melting temperature */
MIN_Tm = 45;                    /* 2. min melting temperature */
MAX_GC = 0.65;                  /* 3. max G+C content */
MIN_GC = 0.45;                  /* 4. min G+C content */
MAX_SINGLE_NT_CONTENT = 0.50;  /* 5. max ratio of single nt */
MAX_CONTIGUOUS_SINGLET_NT = 12;/* 6. max contiguous singlet nt */
MAX_3PRIME_NUM = 4;             /* 7. max contiguous identity at 3' end */
MAX_COMP_CONTIGUOUS_NUM = 5;    /* 8. max contiguous complementarity */
MAX_PAIR_CONTIGUOUS_NUM = 12;  /* 9. max contiguous pair comparison */
MAX_PAIR_SIMILARITY = 0.75; /*10. max allowed similarity for pair probes */
MAX_COMP_SIMILARITY = 0.75; /*11. max allowed similarity for self-complement */
```

In the above, `MAX_SINGLE_NT_CONTENT` is the maximum ratio of the occurrence of any single nucleotide letter—*i.e.*, A, C, G, or T—to the length of the probe. `MAX_CONTIGUOUS_SINGLET_NT` is the maximum length of the consecutive occurrence of any single nucleotide letter in the probe. `MAX_3PRIME_NUM` is the maximum length of complementary base pairs consecutively between a given probe's 3' end and some part within the probe. `MAX_COMP_CONTIGUOUS_NUM`

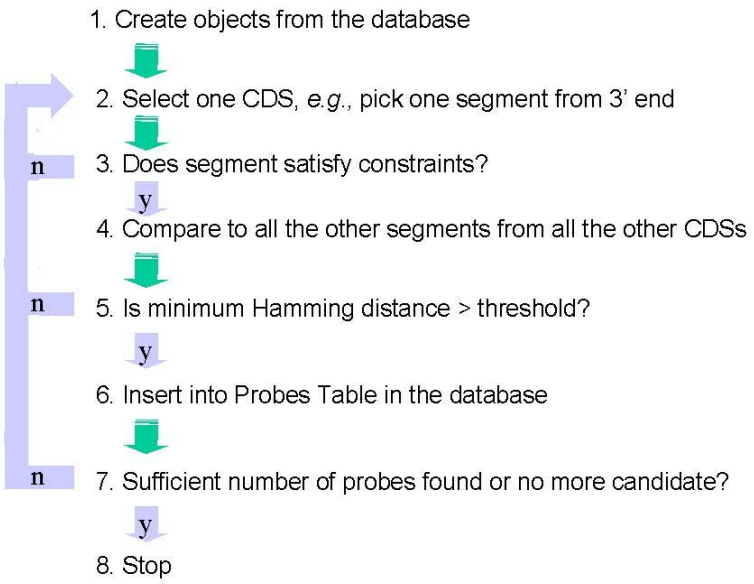

Fig. 2. The main steps of the probe production module of our oligo probe design program.

is the maximum length of complementary base pairs consecutively between any two parts of the probe. `MAX_PAIR_CONTIGUOUS_NUM` is the maximum length of complementary base pais consecutively between a probe and its non-target cDNA. `MAX_PAIR_SIMILARITY` is the maximum total identity, based on Hamming distance, between a probe and its non-target cDNAs.

Note that Constraints 7, 8, and 11 are used to eliminate as many probe candidates that would form internal secondary structures as possible. Although the 3' end checking is not as important in probe design as in primer design, we include it in our program and hope to produce probes with better quality than without this constraint.

4.4. *Probe Production*

Figure 2 describes the probe production process. At the start of the computation and production of putative oligo probes, the probe production module creates all objects described above in the memory from the database. Simply speaking, an array of *Contig* objects is dynamically created from the database first.

Next, the probe production module selects one CDS and tries to pick one segment along the CDS according to a pre-specified picking strategy. Having picked the putative segment, the probe production module needs to do filtering under the given constraints mentioned above.

After a probe passes the process above, the probe production module performs a whole cDNA sequence set comparison using Hamming distance to evaluate the segment. If the Hamming distance between the segment and all other segments from all other CDSs is above the minimal dissimilarity constraint threshold, it is selected as a probe candidate of the CDS and is put into the probes table in the database.

Then the probe production module selects a new CDS and repeats the process above until the number of probes computed is greater than a predefined number—*e.g.*, 10—or there is no more candidate segments to be found.

4.5. *Optimal Probe Set*

When the probe production process terminates, the optimal probe selector module allows the user to select the most optimal set of probes of each ORF according to a definition of what an optimal probe is. In theory, an oligo probe is optimal if the following 3 conditions are satisfied:

- It is as close to the 3' end as possible for optimal cDNA synthesis and labeling of eukaryotic mRNA templates, which utilizes oligo dT primers that bind the 3' polyA tail for transcription initiation.
- It maximizes the Hamming distance between it and all the other CDS to minimize cross-hybridization.
- It has minimal secondary structure so that the sensitivity of hybridization is maximized.

For maximum detection sensitivity of gene expression, the G+C content should be normalized among probes so that the melting temperature is uniform across the probe set. The optimal range is from 45% to 65% in the *S. pombe* genome.

In practice, we use the following rules to select the optimal set of probes for each ORF in the *S. pombe* genome:

- For probes whose distance from the 3' end are less than 500 nucleotide bases, we select those candidates that maximize the Hamming distance;
- For probes whose distances from 3' end are greater than 500 nucleotide bases, we select those candidates that minimize the distance from 3' end.

We see in the next section that the resulting probes selected using these two rules have comparable statistics to the probes designed by some companies we have consulted.

5. Results and Discussions

There are 3 chromosomes in *S. pombe*. We download all 11 contigs in EMBL format (November 2001) from `http://www.sanger.ac.uk/Projects/S_pombe` of the genome sequencing group at the Sanger center. We parse and store all data into our MySQL database at the Genome Institute of Singapore using our program mentioned earlier. Our oligo probe size is 50mer in length. [409] Due to limitations of computing resources, for each ORF excluding duplicates, we produce only about 10 oligo probe candidates for the optimal selection of probes. In order to guarantee the sensitivity of the hybridization and the reliability of experiments, the top two probe candidates are picked for each ORF to construct the whole set of 50mer oligo probes of the *S. pombe* genome.

In our fission yeast functional genomic studies, the initial stage is to design about 10,000 oligos that depict about 5,000 ORFs in total. ORFs under 100 amino acids are excluded by the Genefinder program (C. Wilson, L. Hilyer, and P. Green; unpublished) used for gene prediction. Nevertheless, there are 147 ORFs less than 100 amino acids in length that are included because they either are confirmed experimentally or are reliably predicted with strong significance scores. Currently, there are 4987 annotated ORFs from Sanger Centre (November 2001). From these ORFs, we design about 10,000 oligo probes as described earlier for the Genome Institute of Singapore (January 2002). The probe set is subsequently manufactured by Genset.

To explore whether there are more small ORFs in the genome, we are taking the Matrix-Assisted Laser Desorption Ionization (MALDI) Mass spectrometry approach. The next stage is to include oligos that represent those small polypeptides whose ORFs may not be included in the current annotation.

Some main features of the oligo probes we have produced from the *S. pombe* ORFs described above are as follow:

- A total of 9859 50mer probes are selected for 4929 ORFs (98.84%), after excluding 58 duplicated ORFs.
- The average distance of the probes from the 3' end is 191 nucleotides and the median distance is 141 nucleotides.
- The average cross-homology between the probes is $\leq 61\%$.
- 97.9% of the probes have a distance from the 3' end of ≤ 500 nucleotides.
- 99.9% of the probes have a distance from the 3' end of ≤ 1000 nucleotides.

- 88.1% of the probes have $\leq 60\%$ cross-homology with other probes.
- 98.9% of the probes have $\leq 68\%$ cross-homology with other probes.
- 98.4% of the probes have $45\% \leq$ G+C content $\leq 65\%$.

The availability of genetic sequence information in both public and private databases over the world has shifted genome-base research away from pure sequencing towards functional genomics and genotype-phenotype studies. A powerful and versatile tool for functional genomics is DNA microarray technology which has been vastly applied in monitoring thousands or tens of thousands of genes in parallel and has provided biological insights into gene function and the relevance of the genetic loci for phenotypic traits.

As described earlier, we have developed a program containing several modules that can be used to compute and select optimal oligo probes for an entire genome. It has been successfully applied to the *S. pombe* genome which consists of about 5000 ORFs (in November 2001). The optimal set of oligo probes is composed of two probes from about 10 candidates of each ORF in the genome. From the *in silico* point of view, the quality of the set of probes is very reasonable. However, many microarray experiments across different biological contexts must be analyzed to thoroughly validate the efficacy of the probe designs.

For the improvement of the program that we have developed, the speed of computation of probe candidates should be increased before the program can be used for selecting oligo probes for larger complete genomes.

Acknowledgements

We are grateful to Yuyu Kuang for her useful SQL help; and to Phil Long, Prasanna Kolatkar, Edison Liu, and Mohan Balasubramanian for their invaluable suggestions, comments, and encouragements.