



## Selection of optimal DNA oligos for gene expression arrays

Fugen Li and Gary D. Stormo

Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

Received on January 11, 2001; revised and accepted on May 29, 2001

### ABSTRACT

**Motivation:** High density DNA oligo microarrays are widely used in biomedical research. Selection of optimal DNA oligos that are deposited on the microarrays is critical. Based on sequence information and hybridization free energy, we developed a new algorithm to select optimal short (20–25 bases) or long (50 or 70 bases) oligos from genes or open reading frames (ORFs) and predict their hybridization behavior. Having optimized probes for each gene is valuable for two reasons. By minimizing background hybridization they provide more accurate determinations of true expression levels. Having optimum probes minimizes the number of probes needed per gene, thereby decreasing the cost of each microarray, raising the number of genes on each chip and increasing its usage.

**Results:** In this paper we describe algorithms to optimize the selection of specific probes for each gene in an entire genome. The criteria for truly optimum probes are easily stated but they are not computable at all levels currently. We have developed an heuristic approach that is efficiently computable at all levels and should provide a good approximation to the true optimum set. We have run the program on the complete genomes for several model organisms and deposited the results in a database that is available on-line (<http://ural.wustl.edu/~lif/probe.pl>).

**Availability:** The program is available upon request.

**Contact:** [lif@ural.wustl.edu](mailto:lif@ural.wustl.edu); [stormo@ural.wustl.edu](mailto:stormo@ural.wustl.edu)

### INTRODUCTION

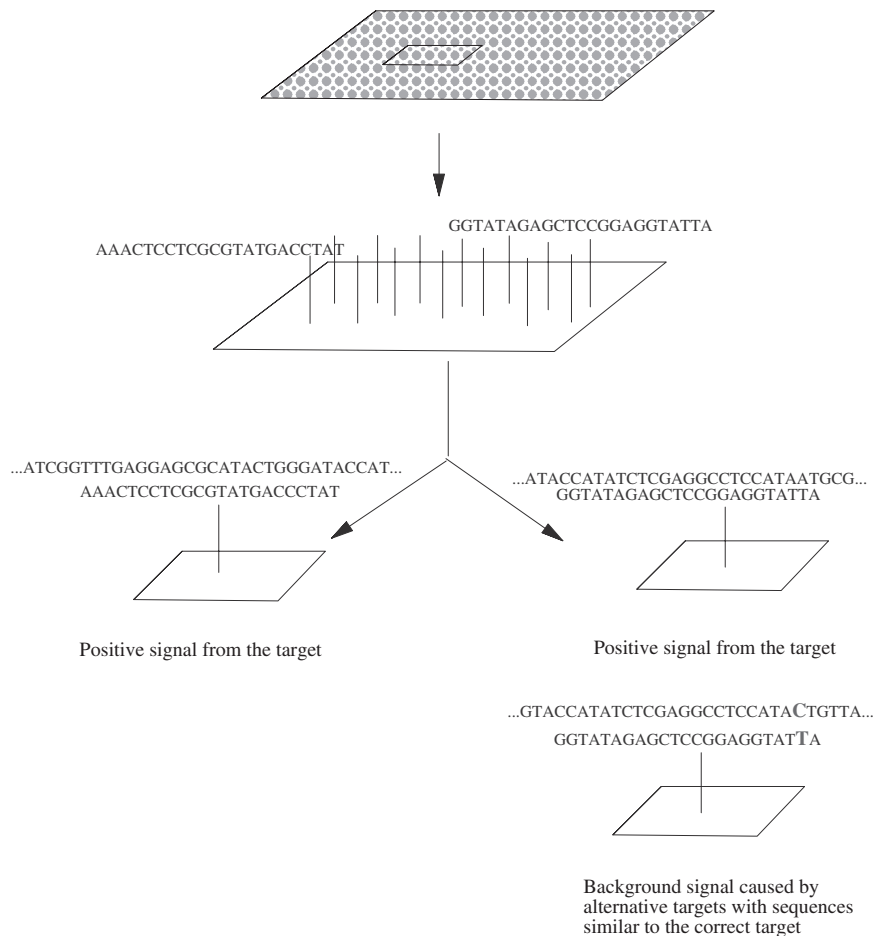
The complete sequences of over 30 microbial genomes and a few eukaryotic genomes are known. In this year, the human genome sequence draft has been finished (International Human Genome Sequencing Consortium, 2000; Venter *et al.*, 2001) and the complete genome sequences of several other metazoans are expected to be completed in the near future. Knowing the sequences of the genes is only the first step in understanding the function of the genome. The intricate circuitry that governs growth, development, homeostasis, behavior and the onset of diseases is largely controlled by the RNA and proteins encoded by the

cognate genes and the complex and dynamic interaction of the genes with the environment. A detailed conceptual view of gene regulatory circuitry in organisms will require extensive expression monitoring at the level of the whole genome (Schena, 1996). The challenge of this biological analysis requires the development and implementation of sophisticated analytical methods. DNA microarray technology offers a great tool for these tasks (Lander, 1999).

### DNA chips

DNA chips are glass surfaces bearing thousands of DNA fragments at discrete sites at which the fragments are available for hybridization. Hybridization of fluorescently labeled RNA and DNA-derived samples to DNA chips allows the monitoring of gene expression or occurrence of polymorphisms in genomic DNA (Gerhold *et al.*, 1999). Two DNA chip formats currently in wide use are the cDNA array format (Schena *et al.*, 1995) and high density synthetic oligonucleotide array format (Lockhart *et al.*, 1996; Pease *et al.*, 1994; Gerhold *et al.*, 1999). Here we focus on oligonucleotide arrays. Oligonucleotide expression arrays include both short oligo (20–25 mers) arrays (Affymetrix geneChip) and long oligo (50–70 mers) arrays (Kane *et al.*, 2000; Bosch *et al.*, 2000).

Affymetrix has married oligonucleotide synthesis and photolithographic computer chip synthesis to generate DNA chips that display 40 000–65 000 DNA oligonucleotides which represent up to 9000 genes on a 1.6 cm<sup>2</sup> glass surface (Gerhold *et al.*, 1999). The oligonucleotides are designed on the basis of sequence information alone. The oligonucleotide approach has some advantages because it allows the user to design probes for each gene to avoid regions that are repetitive or very similar to other known genes (Gerhold *et al.*, 1999; Lipshutz *et al.*, 1999). Although the oligonucleotide approach has advantages, expensive photolithographic technology is not available to most academic laboratories to make customized DNA chips. In the future other methods for arraying oligonucleotides such as ink jet technology (Okamoto *et al.*, 2000) will likely be available to make specifically designed chips rapidly and cheaply. Using



**Fig. 1.** This cartoon shows oligo probes attached on the surface of a chip. Probes can hybridize to their specific targets or alternative targets from a sample pool. The mismatched nucleotides are highlighted.

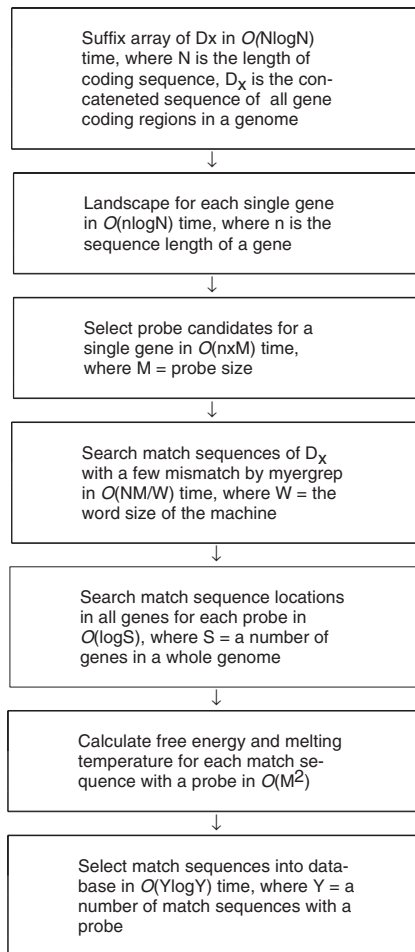
existing spotting array technology, long oligos such as 50 mers or 70 mers have been used to make gene expression arrays (Kane *et al.*, 2000; Bosch *et al.*, 2000). The design of optimum probes would then be of significant benefit. Here we focus on DNA oligo probe design for monitoring mRNA expression.

### The challenge of probe design

Probes on chips can hybridize target mRNA or cDNA from samples. The cartoon in Figure 1 shows how the signals come from the hybridization with targets. If the designed probe in the left part of Figure 1 is unique in the entire coding sequence of a whole genome even with an allowable number of mismatches, the probe has only one target. However, in the right part of Figure 1, if the designed probe is not unique, allowing a few mismatches, and especially if the mismatches are located at the ends of the probe, the probe may have alternative targets which contribute background to the hybridization

signal. Therefore, the quality of data from the oligo chip experiment relies on optimal probes.

The challenge is how to identify the optimum probes for each gene. Empirically, the optimum probe for a gene would be the one with minimum hybridization free energy for that gene (under the appropriate hybridization conditions) and, maximum hybridization free energy for all other genes in the hybridizing pool. Unfortunately, those energies depend on knowledge that is not computable from the sequence alone, at least not currently. For example, the hybridization energy of the probe to the correct gene's cDNA depends on the complete structure of that cDNA. DNA (and RNA) structure prediction programs are not reliable enough for us to accurately predict that structure from the sequence. Another limitation is that the hybridization free energy for every gene, both the correct one and the incorrect ones, depends on the concentration of those genes because this is a bimolecular interaction. For example, an incorrect gene with somewhat higher



**Fig. 2.** The diagram describes the whole procedure of ProbeSelect program.

intrinsic hybridization energy than the correct gene may give a larger signal, i.e. background, than the true gene does simply because it is present in a much greater concentration. We generally do not know in advance the concentration of all the cDNAs; in fact, that is what we are trying to measure. In addition to these issues, which are fundamentally uncomputable at this time, there is also the issue of the complexity of the algorithms, i.e. how the time and memory requirements change when a genome size increases. In principle we could compute the theoretical energy (i.e. ignoring internal structure and concentration) for every potential probe from each gene to every possible hybridization partner, but that task would be computationally intractable.

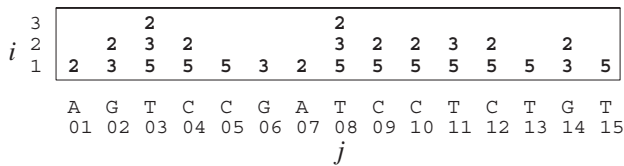
One program exists currently to design oligo probes for DNA chips (Lockhart *et al.*, 1996). This program has been used by Affymetrix to design short oligos (20–25 mers). The algorithm has not been published, but

the probe criteria have been described (Lockhart *et al.*, 1996). Our approach, described in detail below, is to attack the probe design problem in two major steps (composed of several minor steps), which are each tractable and should provide good, although not perfect, predictions of the optimal probes based purely on sequence information. The first major step is to identify, for each gene, the set of candidate probes that maximize the minimum number of mismatches to every other gene in the genome. Since mismatch types and positions affect stability of oligo DNA hybridization, the second major step is to compare those candidate probes more rigorously to determine their hybridization energy, both for their target sequences and for all of the other positions in the genome that match with some allowable number of mismatches. The optimum probes are then picked based on having free energy ( $\Delta G$ ) for the correct target in an acceptable range, and maximizing the difference in free energy to every other mismatched target. The program allows for a number of user selected options, and those are described in the relevant sections.

Our algorithm can be used to design long oligos as well as short oligos. For short oligos, by identifying optimal probes we may be able to reduce the number of probes per gene without a reduction in specificity and sensitivity. This paper describes all of the algorithms used for probe candidate selection and how probes are generated by the program. We have run the program on some model organisms and the selected probe data and gene information are stored in the ChipProbe database which is accessible on-line (<http://ural.wustl.edu/~lif/probe.pl>).

## ALGORITHM

ProbeSelect is written in C++ and was developed on Sun workstations running Solaris. The code is portable for linux and has also been implemented on HP workstations. The program consists of seven major components, described in detail below: (1) make a suffix array of the coding sequences from a whole genome; (2) build a sequence landscape for every gene based on the sequence suffix array; (3) choose probe candidates based on sequence features and the sequence word rank values; (4) search for matching sequences in the whole genome, allowing a certain number of mismatches by the program myersgrep (Myers, 1998); (5) locate match sequence positions in all genes; (6) calculate the free energy ( $\Delta G$ ) and melting temperature ( $T_m$ ) for each valid target sequence; (7) select match sequences that have stable hybridization structures with a probe based on free energy data and allow good discrimination with other targets in the genome. The architecture of the program is shown in Figure 2.



**Fig. 3.** Landscape of a 15-base sequence showing the frequencies of all words within the sequence itself.  $i$  is the word size and  $j$  is the starting position of the word.

### Sequence suffix array

A suffix array (Manber and Myers, 1993) is a sorted list of all the suffixes of a sequence. It takes  $O(N \log N)$  time to build a suffix array, where  $N$  is the length of the sequence. A suffix array permits on-line string searches in time  $O(p + \log N)$ , where  $p$  is the length of the sequence word. Here we build a sequence suffix array of both strands for all coding sequences of a genome, which are concatenated. Alternatively one may choose to use the entire genome sequence, but especially for genomes with lots of non-coding DNA just using the coding regions is probably more appropriate. One could also choose to use only one strand of the coding sequences, the choice depending on how the target sequences are labeled. The details about suffix arrays are described by Manber and Myers (1993).

### Sequence landscape

Detailed information about a landscape is described by Levy *et al.* (1998). A landscape represents the frequency in the database of all the words in the query sequence. In our case the database contains both strands of the coding sequences from the entire genome, but a different set of sequences could be used. A landscape of a simple 15 base query sequence, using the same sequence as a database, shows the basic elements of a landscape (Figure 3). Every cell in the landscape,  $f(j, i)$ , indicates where a word begins  $j$ , and its length  $i$ , and the value of the cell is the frequency of that word in the database. For example, the frequency of 2 at  $f(3, 3)$  indicates that the 3-base word *TCC* starting at position 3 occurs twice in the sequence, the other location being at position 8. All frequency values of 1, indicating a unique occurrence, are not shown in Figure 3. Building a landscape for one gene takes  $O(n \log N)$  time, where  $n$  is the length of the gene coding region and  $N$  is the total length of sequences in the database. In using the suffix array of the concatenated sequences to search for the words, the junction fragments may affect the frequency. However, since our purpose is to search for low frequency words, the effect of junction fragments is relatively small and any that are chosen can be eliminated later. Based on tests of a few model organisms most of the probes selected by

this program, both short and long, are unique even with an allowable number of mismatches.

### Selection of short or long probe candidates

In the results presented here we have used probe sizes from 20 bases to 70 bases, but the length is a parameter set by the user. All potential probes are evaluated by the frequency of matching subwords in the database of other sequences (the entire genome or only other coding regions), as described below. In addition, one can use additional constraints, such as the rules described by Lockhart *et al.* (1996) and used in the Affymetrix probe selection criteria: (1) no single base (As, Ts, Cs or Gs) exceeds 50% of the probe size; (2) the length of any contiguous As and Ts or Cs and Gs region is less than 25% of the probe size; (3)  $(G + C)\%$  is between 40 and 60% of the probe sequence and  $(G + C)\%$  can be adjusted based on  $G + C$  contents of a genome sequence; (4) no 15-long contiguous repeats anywhere in the entire coding sequence of the whole genome (Kane *et al.*, 2000); (5) no self-complementarity within the probe sequence (this constraint is applied in all approaches). The algorithm for self complementarity prediction of DNA primer sequences has been described (Rozen and Skaletsky, 1998) and we modified it for use in the ProbeSelect program to predict self complementarity of each probe sequence. A half matrix is filled by comparing the probe sequence and its complement. For complementary bases the matrix value is 1, otherwise the value is 0. If the maximum value of a diagonal line is more than 30% of the probe length, the probe is substantially self complementary.

A sequence landscape for each gene is used to select probes with low frequency in the rest of the genome. If we were only interested in the frequency in the genome of exact matches to the probe sequence we could simply look it up in the landscape, or we could get a good estimate by multiplying together the frequencies of the sub-words. But we need to know how many approximate matches there are, allowing for some number of mismatches. Fast approximate string matching algorithms, such as *agrep*, are too slow for our needs when we are comparing every potential probe in every gene to all possible target sites in the entire genome. Instead we use the landscape information to get an estimated rank of the number of approximate matches for each potential probe (i.e. sub-word) of each gene. We save the top Q (typically from 10 to 20) probe candidates that are selected for further analysis.

Based on tests of many *Escherichia coli* genes, we have found that summing the frequencies of words at each of several different word lengths (i.e. heights in the landscape) gives a good predictor of the number of matches in the background sequence (i.e. the entire coding sequence or even the entire genome sequence) even with a

few mismatches. Words with the lowest frequencies at all levels of different sub-word lengths are generally the rarest in the rest of the genome, even allowing a few mismatches. Therefore candidate probes are chosen by finding the probe-sized words within a gene sequence that minimize the sum of its sub-word frequencies. For example, if the word  $s$  is comprised of the sequence ATGCCA, ATG is the beginning sub-word and  $f(\text{ATG})$  is the frequency of the word which occurs in coding sequence and its complement in the whole genome. The ‘overlapping frequency’ of word  $s$  is defined to be

$$F(s) = f(\text{ATG}) + f(\text{TGC}) + f(\text{GCC}) + f(\text{CCA})$$

So the general formula for all words in a gene landscape is:

$$F_i(s_j) = \sum_{k=j}^{j+M-i} F_{ik}$$

where  $i$  is column height of the gene landscape,  $j$  is the word position,  $M$  is the probe size and  $j < n - M$ .  $n$  is the sequence length of the gene. For each word size,  $i$ , the ten positions with the lowest overlap frequency are saved. Then all of those positions are ranked by the number of times they occur on those lists. The highest ranking probes are predicted to have the fewest approximate matches elsewhere in the genome. Thorough analyses of small genomes, such as T7 phage, confirms that those predictions are usually quite accurate (data not shown).

### Mismatch searching for candidate probes

The approximate string searching problem is to find all locations at which a query of length  $m$  matches a substring of a text of length  $n$  with  $k$ -or-fewer differences. A fast bit-vector algorithm for approximate string searching has been designed by Myers (1998) to run in  $O(nm/w)$  time where  $w$  is the word size of the machine. Detailed information about this algorithm has been described by Myers (1998). This algorithm is used in the ProbeSelect program to find all match locations for the candidate probes in the coding regions of the genome with four or fewer mismatches (including insertions, deletions and mismatches) allowed for short oligos. We also tried five or fewer, which returns many more locations in the genome, but the extra matches almost never contribute competing sites once the  $T_m$  is considered. The sites with four or fewer mismatches occasionally contribute competing sites with similar  $T_m$ s, so all of those are considered when determining the optimum set of probes. For long oligos, we allow 10 or fewer mismatches for 50 mers and 20 or fewer for 70 mers. Based on a test of 30 *Caenorhabditis elegans* genes, all long oligos selected by this program are unique in the entire coding sequence, even allowing that number of mismatches. We have also used the BLAST

(Altschul *et al.*, 1997) program to compare those probe sequences against the *C. elegans* database to confirm their uniqueness.

### Localization of the match sequences in each gene

The fast approximate string searching is performed on the entire coding and non-coding sequences combined into a single large database. The matches can then be assigned to their exact positions within the specific genes or ORFs. All gene positions can be built into the sorted array. All match positions to the probe can be located in the specific gene by a binary search of the sorted array, which only takes  $O(\log S)$  time where  $S$  is the number of genes or ORFs in the whole genome. If the match sequences are across two genes, they are not valid.

### Free energy and melting temperature calculation

Since different mismatches (including insertion, deletion and mismatch) have different free energy and mismatch locations have different effects on the stability of DNA hybridization, simple counting of mismatches could not determine the stability of DNA hybridization structure. We can not directly calculate free energy and melting temperature between a probe and target on a solid chip, because DNA hybridization behavior on a chip is not the same as that in solution and the parameters on a chip are not available currently. Partial thermodynamic parameters for stacking energy on gel matrix (Kunitzsyn *et al.*, 1996) have an approximate linear relationship with those in solution, although absolute data are different. Therefore, using thermodynamic parameters measured in solution could predict stability of DNA oligo hybridization on chips approximately.

DNA oligonucleotide nearest-neighbor thermodynamic parameters are available (SantaLucia *et al.*, 1996; Allawi and SantaLucia, 1997; SantaLucia, 1998; Allawi and SantaLucia, 1998a,b,c; Peyret *et al.*, 1999) and they allow prediction of oligonucleotide DNA hybridization energies. The parameters for free energy calculation in ProbeSelect are very similar to those in Zuker’s single strand DNA secondary structure prediction program Mfold (Zuker *et al.*, 1999). We could find the minimum energy hybridization structure using a standard dynamic programming approach based on the free energy parameters. However, we already have the alignment of the probe sequence and its target and, since there are at most a few mismatches between them, the lowest energy structure is usually the same as that alignment. Exceptions involve slight rearrangements in, and adjacent to, the mismatched positions. Therefore we designed a fast heuristic to test various alternatives to the alignment. This is  $M$  times faster than the full dynamic programming method where  $M$  is probe size, and nearly always ends up with the same optimum structure. In the few cases where they are

different, the change in free energy is very small.

$T_m$  can be used as a parameter to evaluate probe hybridization behavior. Since it is impossible to know the target DNA concentration, the calculation is approximate, but still useful. With standard conditions,  $T_m$  (Aboul-ela *et al.*, 1985; Rychlik *et al.*, 1990) is calculated for each probe as

$$T_m = \frac{\Delta H}{\Delta S + R * \log(c/4)} - 273.15$$

where  $\Delta H$  and  $\Delta S$  are the enthalpy and entropy for helix formation, respectively, and  $R$  is the molar gas constant ( $1.987 \text{ cal}^\circ\text{C}^{-1} \times \text{mol}$ ).  $c$  is the total molar concentration of the annealing oligonucleotides when oligonucleotides are not self-complementary. As described earlier, this is an intermolecular hybridization so the free energy depends on the concentration. This is not generally known, so we set it to a constant of  $1 \times 10^{-6}\text{M}$ . If concentrations of individual genes were known those could be included in the calculation of  $T_m$ . It is important to note that  $\Delta H$  and  $\Delta S$  are calculated based on standard hybridization solution conditions, because stacking energy parameters for chip conditions are not available currently. If desired, the user can substitute other parameters that are more appropriate for the conditions of hybridization to the chips.

### Selecting matching sequences for the database

Since each probe may have many match regions in the whole genome when allowing four mismatches, it is not necessary to store all match sequences. The hybridization structures between some match sequences and a probe have high free energy and are very unstable. These structures may not be formed under general hybridization conditions, and even if they are formed, they do not contribute significantly to the background. Therefore, these sequences have no effects on the selection of the probe. However, some match sequences have large effects if they hybridize with a probe because they have free energies close to that of the probe to its target. The difference of free energy between the hybridization structures with and without mismatches is used as an index to filter the sequence and other data. Sites with a difference of  $10 \text{ kcal mol}^{-1}$  or less in free energy usually have a difference in  $T_m$  of less than  $20^\circ\text{C}$  and can contribute to background hybridization. Therefore all of those mismatch sites are also stored in the database.

All sequences matched with a probe are sorted into an array based on free energy by Quicksort which takes  $O(Y \log Y)$  time where  $Y$  is the number of sequences. An output filter checks the difference of free energy between all sequences and the probe to determine whether this sequence should be stored in the database or not. This algorithm takes  $O(Y)$  time in the worst case.

## RESULTS

### Performance on a few organisms for selection of short oligos

ProbeSelect was tested on a few organisms to generate probes for each gene. On a Sun workstation it took under 2 minutes to finish T7 phage genome, which is 39 937 bps and includes 60 genes. About ten probes of 20 bases were generated for each gene and match sequences to each probe were searched with four mismatches. However, it took about 1.5 days to finish the *E. coli* genome which is about  $4.6 \times 10^6$  bps and includes about 4300 genes, and almost 4 days to finish *Saccharomyces cerevisiae* genome which is about  $1.2 \times 10^7$  bps and includes about 6000 genes. The size of a probe for *E. coli* is 23 bases and for *S. cerevisiae* is 24 bases and the match sequences were searched with four mismatches.

### Short probes generated by ProbeSelect

Tables 1 and 2 show the ten probes selected for the yeast CDC28 gene with and without, respectively, the criteria used in Affymetrix probe selection based on overall base content, as described above. The number in the 'num-matches' column is how many matches there are in the entire yeast coding sequence allowing up to four mismatches with the probe. A '1' in that column indicates that the correct target for the probe is the only match; numbers greater than one indicate that there are other genes with matches to the probe, allowing a few mismatches. The probes selected without the pre-screen of nucleotide content (Table 2) may have high GC or AT content depending on nucleotide content and distribution of the genome. Most of the probes selected by the two approaches are the same, or shifted by only one or two bases. A few of the probes in Table 2 have high GC content because the yeast genome has high AT content (64%) and a high GC sequence is more likely to be a unique region. However, all probes generated in Table 1 have about 50% GC content, which is a common criterion for PCR primers and probes.

Three model organisms have been tested. The results showed that all probes from T7 phage are unique with four or fewer mismatches. More than 60% of probes from *E. coli* are unique with four or fewer mismatches when the probe size is increased to 23 bases, and more than 80% of probes from *S. cerevisiae* are unique with four or fewer mismatches when the probe size is 24 bases. Obviously, increasing the probe size makes it much easier to find unique probes for each gene in a large genome. Determination of probe size and a number of mismatches for one of the genomes will help design good probes to limit nonspecific hybridization on a DNA array although the short probe size should be in a range from 20 to 25 bases.

**Table 1.** Probes for yeast *cdc28* gene generated by ProbeSelect\*

gene-name	probe-name	probe-sequence	b-pos	num-matches
CDC28	PROBE-1	AAACTCCTCGCGTATGACCCTATT	820	1
CDC28	PROBE-2	TCCTCGCGTATGACCCTATTAACC	824	1
CDC28	PROBE-3	CCAAGTCTAGATCCACGCGGTATT	784	1
CDC28	PROBE-4	CGCGTATGACCCTATTAACCGGAT	828	3
CDC28	PROBE-5	TGCATACTGCCACTCACACCGTAT	372	1
CDC28	PROBE-6	CCTATTAACCGGATTAGCGCCAGA	838	1
CDC28	PROBE-7	TACTGCCACTCACACCGTATTCTG	376	1
CDC28	PROBE-8	CTATGGTATAGAGCTCCGGAGGTA	523	2
CDC28	PROBE-9	GGTATAGAGCTCCGGAGGTATTAC	527	1
CDC28	PROBE-10	AGTATTGGGAACGCCGAATGAAGC	678	3

\*Probe candidates are selected with the base composition criteria. b-pos means the beginning position of a probe in a gene. num-matches means how many of positions a probe can match an entire coding sequence with 4 or fewer mismatches.

**Table 2.** Probes for yeast *cdc28* gene generated by ProbeSelect\*

gene-name	probe-name	probe-sequence	b-pos	num-matches
CDC28	PROBE-1	AAACTCCTCGCGTATGACCCTATT	820	1
CDC28	PROBE-2	CTCGCGTATGACCCTATTAACCGG	826	1
CDC28	PROBE-3	AGACGAGGGTGTCCAGTACAGC	141	1
CDC28	PROBE-4	TGCATACTGCCACTCACACCGTAT	372	1
CDC28	PROBE-5	CAAGTCTAGATCCACGCGGTATTG	785	2
CDC28	PROBE-6	GAGCAGCCATCCACCCCTACTTCC	863	1
CDC28	PROBE-7	GAGCTCCGGAGGTATTACTGGGTG	533	1
CDC28	PROBE-8	TATTAACCGGATTAGCGCCAGAAG	840	1
CDC28	PROBE-9	TACTGCCACTCACACCGTATTCTG	376	1
CDC28	PROBE-10	TGGTATAGAGCTCCGGAGGTATTA	526	2

\*Probe candidates are selected without the base composition criteria. b-pos means the probe starts at the position of the gene. num-matches means how many of positions a probe can match an entire coding sequence with 4 or fewer mismatches.

We tested how well programs to select PCR primers, of which there are several, would work to select unique probes for hybridization. PCR primer selection is less stringent than our task because in order to get an inappropriate PCR product both primers must hybridize within a relatively short distance of each other. But for microarrays, an inappropriate hybridization anywhere in the entire coding sequence of a genome can lead to high background. We tested one PCR primer selection program (<http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer>) to see how its selections compare to those generated by our program. Table 3 lists the selected primers for three yeast genes (other genes give similar results) and the number of matching sites in an entire coding sequence of a whole genome if four mismatches are allowed. In general each of them has multiple potential target sites in the entire coding sequence, whereas our program returned primarily sites that are unique, and in general have many fewer alternative hybridization sites. Although the program is adequate for identifying good

primers for PCR, it would not work nearly as well for identifying optimum probes for microarrays.

### Probe hybridization behavior

Free energy parameters are used to predict the minimum energy hybridization structure between a probe and its matching sequence. In Table 4,  $\Delta G$  and  $T_m$  for hybridization structures for the yeast gene CDC28 probes (from Table 1) are listed. Only probe-8 has an alternative target in the genome with a free energy difference within  $-10 \text{ kcal mol}^{-1}$  compared to the true target. Based on the data in Table 4 one could pick the optimal probe (or small set of probes) for the CDC28 gene.

The algorithm for hybridization free energy calculation has been compared with the standard dynamic programming algorithm with free energy rules to fill the matrix. 95% of the free energy calculations and hybridization structure predictions are identical. Only 5% of them have minor differences which do not affect the probe selection. The algorithm which is included in ProbeSelect takes  $O(M^2)$  time, but the standard dynamic programming

**Table 3.** Primers for three yeast genes generated by web-primer\*

gene-name	primer-type	primer-sequence	num-matches
CDC28	forward	GAGCGGTGAATTAGCAAATACAA	2
CDC28	reverse	GATTCTTGGGAAGTAGGGGTGGATG	2
CYS3	forward	CAAGAATCTGATAAAATTTGCTACC	6
CYS3	reverse	GTTGGTGGCTTGTTC AAGGCTTG	2
ACT1	forward	GGATTCTGAGGTTGCTGCTTTGGT	1
ACT1	reverse	ACTTGTGGTGAACGATAGATGAC	3

\*The URL of web-printer program is <http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer>. Only two best primers for each gene are shown here. num-matches means how many of positions a primer can match an entire coding sequence with 4 or fewer mismatches.

**Table 4.** Detail information of all probes for yeast *cdc28* gene\*

gene-name	probe-name	probe-match-seq alignment	m-gene	mb-pos	mis	$\Delta G$	$\Delta\Delta G$	$T_m$	h-strand
CDC28	probe-1	AAACTCCTCGCGTATGACCTATT TTGAGGAGCGCATACTGGGATAA	CDC28	820	0	-32.0	0	79.3	noncode
CDC28	probe-2	TCCTCGCGTATGACCCTATTAACC AGGAGCGCATACTGGGATAATTGG	CDC28	824	0	-31.9	0	79.7	noncode
CDC28	probe-3	CCAAGTCTAGATCCACGCGGTATT GGTTCAGATCTAGGTGCGCCATAA	CDC28	784	0	-32.2	0	80.1	noncode
CDC28	probe-4	CGCGTATGACCCTATTAACCGGAT GCGCATACTGGGATAATTGGCCTA	CDC28	828	0	-32.3	0	80.3	noncode
CDC28	probe-5	TGCATACTGCCACTCACACCGTAT ACGTATGACGGTGAGTGTGGCATA	CDC28	372	0	-33.0	0	82.9	noncode
CDC28	probe6	CCTATTAACCGGATTAGCGCCAGA GGATAATTGGCCTAATCGCGGTCT	CDC28	838	0	-32.0	0	81.0	noncode
CDC28	probe-7	TACTGCCACTCACACCGTATTCTG ATGACGGTGAGTGTGGCATAAAGAC	CDC28	376	0	-32.0	0	81.1	noncode
CDC28	probe-8	CTATGGTATAGAGCTCCGGAGGTA GATACCATATCTCGAGGCCTCCAT	CDC28	523	0	-30.5	0	79.5	noncode
CDC28	probe-8	CTATGGTATAGAGCTCCGGAGGTA TCTACCATATCTCGAGGCCTTTAT	SLT2	586	4	-22.5	-8.0	69.1	noncode
CDC28	probe-9	GGTATAGAGCTCCGGAGGATTAC CCATATCTCGAGGCCTCCATAATG	CDC28	527	0	-30.2	0	78.7	noncode
CDC28	probe-10	AGTATTGGGAACGCCGAATGAAGC TCATAACCCTTGGCGCTTACTTCG	CDC28	678	0	-32.9	0	81.9	noncode

\*m-gen means a gene name of a sequence a probe match. mb-pos is the starting position of the gene that a probe match. mis is the number of mismatches when a probe matches the sequence. h-strand means what sequence strand a probe hybridizes with.

algorithm with free energy rules takes  $O(M^3)$  time.

### Long probe selection

Long oligo arrays have advantages over both short oligo gene chips and cDNA microarrays (Kane *et al.*, 2000; Bosch *et al.*, 2000). We tested the program to select 50 mer and 70 mer oligos for several model organisms such as *E. coli*, yeast and *C. elegans*. For yeast, it takes 92 h to select five 50 mer oligos for each gene or ORF and 96 h to select five 70 mer oligos for each gene or ORF. The results indicated that the running time for long oligos is not significantly longer than that for short oligos. In Table 5, 70 mer probes for ten *C. elegans* genes are shown. All of the probes in Table 5 are unique in the entire *C. elegans* coding sequence even with 20 mismatches allowed. Those probes are also confirmed to have no highly similar regions by BLAST searching against the *C. elegans* genomic database. GC and AT nucleotides are evenly distributed across all probe regions. Similar results are observed for 50 mer oligos.

### DISCUSSION

The short probe size is usually 25 bases on the DNA chips made by Affymetrix. However, the actual size depends on the probe hybridization stability and genomic size of the organism. For practical reasons, 20 bases are minimal size for the probe. All probes selected for T7 phage by ProbeSelect do not have any other match in the entire coding sequence even if four mismatches are allowed. However, for *E. coli*, most probes with 20 bases have more than 30 matches in the entire coding sequence if four mismatches are allowed, and about 70 matches in the entire coding sequence if five mismatches are allowed. The probe size should increase as the genome size becomes larger. For yeast, most probes of 20 bases have about 70 matches in the entire coding sequence if four mismatches are allowed. However, when probe size increases to 23 bases for *E. coli* more than 60% of the probes are unique even if four mismatches are allowed. Thus, the short probe size has to be determined for each organism.

**Table 5.** Long probes with 70 mers for 10 *C. elegans* genes generated by ProbeSelect\*

gene-name	probe-sequence	b-pos	num-matches
B0511.10	GATGCTAATCCATTTCGACACGGAATGGGAAGTTGTTACGC- AACGTGCGTGGCTTATGCACTGGGCACTTT	589	1
C01F6.4	TTCCCGTCTACAGTTACAAATTTGCGCGTTTGAATCTCTC- GCTACTCACCTCTAGATGTATCTGGTTAT	975	1
C27A2.3	ATCGACACTGCAAACCGATGTGAAACATTTCGATTCGCTAG- CTGCTGACATTGAGGACGATATGCTTAATT	316	1
F07A5.7	AGAGACAACCTTGCTGAGTCCGAGTCCGTAACAATGCAAAA- CCTCCAGAGAGTCCGCAGATACCAACACGA	2075	1
F54C9.8	TGGACTCATGACATCAGTGGCAGCAACTGCTTCCGACTT- TCATCGAACGAGTTCCGCTAACTACGTCGTT	1026	1
K12F2.1	TCCGTGGAAAGCTTAAGCTTAGCAACGACATCACTTACTA- TCACTTCTGCTCGCAGGCCGAGCTTACCAT	968	1
R107.8	GAATACGCTCGGTACATACATTTGCGTGTGTCCCCAAGGA- TTCCTTCCCTCCGATTGTCTAAAACCTGGG	510	1
T05E11.4	CCAATATACGGACTGTTTGACGCTGATCCGCATGGTATTG- AAATCTACTTGACTTACAAATATGGACCTA	745	1
Y73B6A.5	CTTTGGCCGCACCTATGAAGGTGAACGTCCTTTGGCTCAAC- AACCTAATACGATTGCGTGTGCTAGTCTT	2070	1
ZK381.1	CGAACAATAGATAAACTGTTGGTCACCGCATTCCAGAGCA- GTGTTCTGCCGCAAAGCGTGTGTCTTCAA	193	1

\*b-pos means the probe sequence starts at the position of the gene. num-matches means how many of positions a probe can match an entire coding sequence with 20 or fewer mismatches.

Setting the number of mismatches is very important for searching similar regions with each probe. A few genes have been tested for *E. coli*, when the probe size is 23 bases. Free energies ( $\Delta G$ ) of all match regions with the probes are calculated. The difference in free energy ( $\Delta\Delta G$ ) between the hybridization structure with perfect match and that with mismatches is also calculated. The  $\Delta\Delta G$  values from 0 to  $-10$  for four or less mismatches contain all of the  $\Delta\Delta G$  values for five or less mismatches for all tested genes. This means all hybridization structures with four or less mismatches have enough information to select probes with 23 bases for *E. coli*. We have also tested a few yeast genes and selected 24 bases as the probe size and four for the number of mismatches. Thus, for different genomes, we may have different sizes for a probe and a different number of mismatches allowed. Choosing the reasonable probe size and the number of mismatches can save significant computation time and computer memory, which is critical for a large genome. Others have also suggested that four mismatches are the maximum number allowed for probe size under 25 bases (Wodicka *et al.*, 1997). For longer oligos many more mismatches are needed to identify potential cross-hybridizing genes. We have 10 and 20 for 50 mer and 70 mer probes, respectively, but it may also work just as well to search for similar regions using BLAST.

It takes four days for ProbeSelect to finish the yeast genome when on average eight probes with 24 bases are selected for each gene and four mismatches are allowed to search similar regions to each probe. It takes a few

weeks to finish the worm genome. Thus, it is unlikely for ProbeSelect to run the human genome in reasonable time if the algorithm is not improved. Based on the algorithm used in ProbeSelect, we could improve it to change the performance for the program. The myersgrep algorithm in ProbeSelect for approximate matching takes about 50% of the total time. Since a suffix array for the coding sequence has been built for the gene landscapes, it could be used for approximate string search which should be much faster than myersgrep. Work on optimizing that approach is currently underway. It is also possible to parallelize the algorithm to run on multiple processors and reduce the time required accordingly.

ProbeSelect aims to allow the selection of optimum oligo probes for each gene, given the entire genomic sequence. The user has a number of options that affect the criteria of optimality: the length of the probe sequence; whether or not to imposed probe composition constraints; what hybridization energy parameters to use. By default the program uses a fast approximate matching search algorithm (myersgrep) to identify alternative targets in the genome, but that can be substituted with methods such as BLAST, if desired. In addition, one can rank the selected probes by various criteria:  $\Delta\Delta G$ ,  $T_m$ , or simply the number of mismatches with alternative targets. Although not explored in this paper, a number of other constraints and criteria might be applied. For example, one might want to only select probes from the 3' ends of genes because those are more heavily labeled in oligo-dT primed labeling reactions. This could be enforced by

only providing those regions of the genes for analysis by the program, or that could simply be part of the ranking procedure from which the final probe set is selected. Since the hybridization free energy depends on the target sequence concentration, one could estimate that based on measures of codon bias. That might provide additional useful information that would allow one to avoid, in particular, alternative targets that are expected to be highly expressed, and therefore contribute significant background signal.

## ACKNOWLEDGEMENTS

This work reported here has been supported by a grant (ER61606) from the Department of Energy. We thank Jon Sauer for the use of his HP computers for testing various parts of the program. We also thank Gene Myers and Sam Levy for discussions. We are very grateful to the referees for their valuable comments on the manuscript.

## REFERENCES

- Aboul-ela, F., Hoh, D. and Tinoco, I.J. (1985) Base-base mismatches—thermodynamics of double helix formation for dCA3XA3G + dCT3YT3G (X, Y = A, C, G, T). *Nucleic Acids Res.*, **13**, 4811–4825.
- Allawi, H.T. and SantaLucia, J.J. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry*, **36**, 10581–10594.
- Allawi, H.T. and SantaLucia, J.J. (1998a) Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry*, **37**, 2170–2179.
- Allawi, H.T. and SantaLucia, J.J. (1998b) Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Res.*, **26**, 2694–2701.
- Allawi, H.T. and SantaLucia, J.J. (1998c) Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, **37**, 9435–9444.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped blast and psi-blast: a new generation of protein database search program. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bosch, J.T., Seidel, C., H.Lam, S.B., Tuason, N., Saljoughi, S. and Saul, R. (2000) Validation of sequence-optimized 70-base oligonucleotides for use on DNA microarrays. In *The TIGR Genome Sequencing and Analysis Conference*. Poster, Maimi Beach, Florida.
- Gerhold, D., Rushmore, T. and Caskey, C.T. (1999) DNA chips: promising toys have become powerful tools. *Trends Biochem. Sci.*, **24**, 168–173.
- International Human Genome Sequencing Consortium, (2000) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Kane, M.D., Jatke, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
- Kunitsyn, A., Kochetkova, S. and Florentiev, V. (1996) Partial thermodynamic parameters for prediction stability and washing behavior of DNA duplexes immobilized on gel matrix. *J. Biomol. Struct. Dyn.*, **14**, 239–244.
- Lander, E.S. (1999) Array of hope. *Nature Genet.*, **21**, 3–4.
- Levy, S., Compagnoni, L., Myers, E.W. and Stormo, G.D. (1998) Xlandscape: the graphical display of word frequencies in sequences. *Bioinformatics*, **14**, 74–80.
- Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet.*, **21**, 20–24.
- Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- Manber, U. and Myers, E.W. (1993) Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, **22**, 935–948.
- Myers, E.W. (1998) A fast Bit-vector algorithm for approximate string matching based on dynamic programming. In *Ninth Combinatorial Pattern Matching Conference*. Piscataway, NJ, pp. 1–13.
- Okamoto, T., Suzuki, T. and Yamamoto, N. (2000) Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat. Biotechnol.*, **18**, 438–441.
- Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl Acad. Sci. USA*, **91**, 5022–5026.
- Peyret, N., Senerviratne, P.A., Allawi, H.T. and SantaLucia, J.J. (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G and T.T mismatches. *Biochemistry*, **38**, 3468–3477.
- Rozen, R. and Skaletsky, H.J. (1998) [http://www-genome.wi.mit.edu/genome\\_software/other/primer3.html](http://www-genome.wi.mit.edu/genome_software/other/primer3.html).
- Rychlik, W., Spencer, W.J. and Rhoads, R.E. (1990) Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.*, **18**, 6409–6412.
- SantaLucia, J.J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
- SantaLucia, J.J., Allawi, H.T. and Senerviratne, P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
- Schena, M. (1996) Genome analysis with gene expression microarrays. *BioEssays*, **18**, 427–431.
- Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
- Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
- Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. and Lockhart, D.J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, **15**, 1359–1367.
- Zuker, M., Mathews, D.H. and Turner, D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction. In Barciszewski, J., Clark, B.F.C. and Clark, (eds), *A Practical Guide in RNA Biochemistry and Biotechnology*. Kluwer Academic NATO ASI Series edition, Dordrecht.