African Journal of Biotechnology Vol. 6 (13), pp. 1582-1586, 4 July 2007
Available online at http://www.academicjournals.org/AJB
ISSN 1684 - 5315 © 2007 Academic Journals

Full Length Research Paper

A comparative analysis of existing oligonucleotides selection algorithms for microarray technology

Ezekiel F. Adebiyi

Department of Computer and Information Sciences, Covenant University P. M. B 1023, Ota, Nigeria. E-mail: eadebiyi@sdsc.edu.

Accepted 18 May, 2007

In system biology, DNA microarray technology is an indispensable tool for the biological analysis involved at the level of the whole genome. Among the sophisticated analytical problems in microarray technology at the front and back ends, respectively, are the selection of optimal DNA oligonucleotides (henceforth oligos) and computational analysis of the genes expression data. A computational comparative analysis of the methods used to select oligos is important since the design and quality of the microarray probes are of critical importance for the hybridization experiments as well as subsequent analysis of the data. In an attempt to enhance efficient and effective design at the front end, a computational comparative analysis was performed on oligos selection tools using the barley ESTs, as well as the *Saccharomyces cerevisiae, Encephalitozoon cuniculi* and human genomes. The analysis also shows that a large number of the existing tools are difficult to install and configure. For cross hybridization test, most rely on BLAST and therefore design ill specific oligonucleotides. Furthermore, most are non-intuitive to use and lack important oligo design and software features.

Key words: System biology, microarray technology, oligo, genome, coding sequence, expressed sequence tag (EST).

INTRODUCTION

The sequencing race has ended and the functional race has already begun. The expression “system biology” (Kitano, 2002) is used today mostly to describe attempts at unraveling molecular systems (the function of the genome), beyond the traditional level of single gene and single protein, focusing on the intricate circuitry that governs growth, development, homeostasis, behaviors and the onset of diseases, which 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). Microarray technology enables simultaneous gene expression analysis of thousands of genes, enabling a snapshot of an organisms’ transcriptome at an unprecedented resolution. The close correlation between gene transcription and function, allows the inference of biological processes from the assessed transcriptome profile. Among the sophisticated analytical problems in microarray technology at the front and back ends respectively, are the selection of optimal DNA oligos and computational analysis of the genes expression. Three basic procedures are involved in the setup of a microarray experiment. These include design of the DNA chips (arrays), gene expression profiling experiments and computational analysis of the array data (Baldi and Hatfield, 2002).

DNA chips are glass surface 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 poly-morphisms in genomic DNA (Schena, 1996). Two DNA chips format currently in wide use are, the cDNA array format and high density synthetic oligonucleotide array format (oligos). The oligonucleotide approach has some advantages because it allows the user to design oligos for each gene to avoid regions that are repetitive or very similar to other genes (Gerhold et al., 1999; Lipshutz et al., 1999). Furthermore, spotting pre-synthesized oligonucleotide has many advantages, such as high sensitivity, convenience, cost effectiveness andappreciable correctness when compared to the cDNA array format. Oligonucleotide expression arrays include both short oligo (20-25 mers) arrays (Affymetrix gene Chip) and long oligo (50 - 70 mers) arrays. It has been noted by Le Roch et al. (2003) that 25-mer oligos could be hybridization problematic and longer oligos might be needed to generate accurate expression profiles.

The usage of oligonucleotide in this functional race is amazing. For example, an oligonucleotide-based antiviral compound with potential to serve as a prophylactic and therapeutic agent in case of an influenza is been built. This compound has also the potential to be used as a viable drug against dreadful resistant viruses (like HIV) and others certain viral infections like hepatitis C and tuberculosis. To show more light on this application, one of the lead compound is a 40 mer fully degenerate phosphorothioate oligo and it targets the common chemical and structural properties of specific motifs found in most enveloped viruses that are required for the fusion of the viral membrane with the host cell membrane, thereby preventing entry of the virus into the cell.

In this paper, existing tools for oligos selection are computational compared as regard effectiveness (specificity) and efficiency (computer speed and memory requirement). It is important to note that a computational comparative analysis of the methods used to select oligos is important since the design and quality of the microarray probes are of critical importance for the hybridization experiments (microarray experiment middle end phase) as well as subsequent analysis of the data (microarray experiment back end phase).

MATERIALS AND METHODS

Four data sets are used in this study. The first is the barley ESTs. Nearly 300,000 publicly available ESTs derived from barley cDNA libraries are currently present in dbEST. These sequences have been quality-trimmed, cleaned of vector and other contaminating sequences (such as repeats), pre-clustered and clustered into final assemblies of "contigs" (that is, overlapping EST sequences) and "singletons" (that is, non-overlapping EST sequences) called unigenes. This barley dataset was obtained from http://har-vest.ucr.edu/ using the HarvEST viewer version 1.45. This viewer downloaded 29 Jan., 2006 has a collection of 26,634 contigs and 26,606 singletons forming 53,240 unigenes of a total of 43,464,144 bases. The second dataset used in this work is the baker yeast genome. The baker yeast, *Saccharomyces cerevisiae* is one of the well studied organisms. The yeast genome consists of 6343 coding region sequences totaling 9.5 MB in nucleotides size. The other two are the *Encephalitozoon cuniculi* and human genomes. They are used in our attempt to select oligos from biological samples which consist of a mixture of pathogen transcripts, the *E. cuniculi* and host cell transcript, the Human. The genome of this pathogen is presently known as the smallest eukaryotic with a size of 2.9 MB and represents an increasing danger to human health.

For specificity test, the tools are tested using Yeast, *E. cuniculi* and the human genomes. And for efficiency test, they are tested on the ESTs of barley.

RESULTS AND DISCUSSION

Efficiency consideration

Recent overviews listed 12 (Nordberg, 2005), 9 (Sebastien et al., 2005) and 13 (Combes et al., 2005) existing oligo selection programs, but a comprehensive search shows that we have 29 existing oligo selection programs (26 for annotated genomics and 3 for EST sequences with non-annotated genomes).

A new categorization of these tools has been previously presented (Adebiyi, 2007) Existing tools for oligos selection were categorized along the techniques employed to get oligos candidates. The methods developed so far can therefore be grouped into two classes: The first class use some pattern matching technique(s) to select possible oligo candidates while the other classes simply use all l-mer possible from the target sequences, where l is the length of the oligos desired. The advantage of the first approach is that checking specificity of each oligo candidate will not be necessary again at the filtering stage. A complete listing of all programs till date in these two categorizes can be found at http://www.covenantuniversity.com/bioinformatics/oligotools/ExistingOligo1.htm. A tool is named after its author(s), if the program has no name. For example, recent programs in the second class include YODA (Nordberg, 2005) and SEPON (Henrik et al., 2004) while ProbeSelect (Li and Stormo, 2003) and OligoSpawn (Zheng et al., 2004) have been developed using the first approach.

Most oligos selection programs (including SEPON) rely heavily on an external program, e.g. BLAST. Note that BLAST may select oligos which are greater than 86% identical to non-target sequences, regardless of similarity threshold used (Nordberg, 2005). This is one of the shortcomings tackled in YODA in the development of a customized sequence similarity search tool, named SeqMatch. SeqMatch is said to be fast (efficient) for identity greater than 80% but for identity less than 80%, the algorithm requires significantly more time. For example, on the barley ESTs dataset, for oligo length (henceforth l) = 33, for identity less than 80%, YODA runs for about 9 days (12785.68 min) and for l = 70, it runs for about 5 days (6845.43 min) on a 3.0 Ghz Pentium IV PC with 512MB RAM, while for identity greater than 80%, it ran for more than 2 days (3221.30 min), before the program was stopped. So it is un-suitable for large ESTs, such as the barley EST datasets. This inefficiency observed in YODA seem to be an inherent characteristic of all the algorithms designed using the based-technique of the second approach, in essence, the use of all l-mers possible from the target sequences as candidate oligos.

ProbeSelect selects candidate oligos using landscape (Li and Stormo, 2003). However, it requires respectively 1.5 days for 4.6 MB dataset and 4 days for 12 MB dataset on a SUN Workstation. SEPON is designed for ESTs dataset but difficult to configure and install. Several attempts to install SEPON failed. The contacting author, Henrik Hornshoj, could not tell why SEPON will not install on our system before this manuscript was submitted. OligoSpawn select candidate oligos via a two-phase algorithm, using the notion seeds (a substring of lengthq), which are then extended to candidate oligos. OligoSpawn has been carefully engineered to efficiently identify all unique oligos in the ESTs (to find short oligos of length 33 (q = 11), it ran on a previous Barley dataset of 28 MB for 2 h and 26 min using a 1.2 Ghz AMD machine), but it was observed to be very inefficient for finding unique oligos, when q ≥ 13. This is because the run-time of the algorithm is exponential in q. For example, the experiment performed with OligoSpawn for which oligos of length l = 33 (q=11) were selected may have better results with longer oligos of l = 39 (q = 13), since the additional six nucleotides are not too long to form a stem loop or self hybridize, but even increase stringency of binding (Esiobu, 2006) (increased specificity), which in turn helps reduce the possibility of false binding. For shorter oligos, e.g. l = 20 to l = 25 or l = 33 bases, Hamming distance, d =4 or d = 5 mismatches have been used to identify potential cross-hybridizing genes. And for longer oligos, e.g. l = 50 bases and l = 70 bases, Hamming distance, d = 10 and d = 20 mismatches respectively has been used (Li and Stormo, 2003; Zheng, et al., 2004). Furthermore, for large datasets (like the barley ESTs dataset in this work), OligoSpawn tractable (for q=11) quadratic run time becomes inefficient as its runs for about 6 days to find oligos of length l = 33 using a 3.0 Ghz Pentium IV machine.

The author's (Adebiyi, 2007) contributions are in two folds. First, the work presented a novel algorithm to re-implement the two-phase of OligoSpawn using Suffix tree, cleverly, engineering the resulting algorithm to be efficiently suitable for identifying both short and long unique oligos. The validity of the resultant algorithm has been checked by benchmarking it against OligoSpawn. Experimentation with the algorithm shows that the time for finding long oligos is insignificant to the time used for selecting short oligos. This is also reported about ProbeSelect (Li and Stormo, 2003). For example, on the barley dataset, the algorithm runs for 5142.6 min to find oligos of size l = 39 (q = 13). On the same dataset for l = 50 (q = 8), l =60 (q = 10) and l = 70 (q = 6), experimentations show that these oligos can be computed in the neighborhood of 4 days. Second, the recent algorithm designed for selecting oligos for ESTs, SEPON, rely on external programs like Bioperl, perl MLDBM, BLAST (for annotation and specificity testing), MELTING (Novere, 2001) and mfold (Zuker et al., 1999). In the same spirit of the work behind the development of YODA (Yet-Another Oligonucleotide Design Algorithm), the work also reduces this dependence using suffix tree and other simple pattern matching techniques (for checking potential stem-loop structures, identifying prohibited sequences, and dimerization potential) to external programs that only include BLAST (for annotation) and MELTING.

Effectiveness consideration

In designing oligos, three important filtering criteria are important, namely, sensitivity and specificity of the individual probes and its consistency among the set of probes (Nordberg, 2005). The most important point among these criteria is the specificity of the oligonucleotides. And this refers to the inability of the probe to bind strongly to non-target sequences that may be present during the hybridization. This can be implemented using the following criteria for checking excessive sequence similarity (Kane et al., 2000; Hughes, et al., 2001):

1.) The oligo sequence must not have more than > 75-80% of similarity with a non-target sequence present in the hybridization pool.

2.) The oligo sequence must not include a stretch of identical sequence > 15 contiguous bases.

The specificity of tools is discussed in the discussion that follows, using a naive implementation of the above criteria in C programming language. Note that Nordberg procedure for checking the above specificity criteria guarantee an optimal effectiveness. Specificity checking has mostly been implemented using BLAST. Note that BLAST may select oligos which are > 86% identical to a non-target sequences, regardless of similarity threshold used (Nordberg, 2005).

In an overview for a complex target mixture, where targets are extracted from biological samples which consist of a mixture of pathogen transcripts and host cell transcript, Rimour et al. (2005) showed that classical approaches, the ones we discussed above, failed to find specific 50 mer oligonucleotides for approximately 60% of the CDSs. In the case they considered, the target is the *E. cuniculi* and the mixture of pathogen transcripts and host cell transcript is the union of *E. cuniculi* genome and Human genome. Five tools, namely, YODA, Probe-sel, PICKY, Array Designer 4.0 (a commercial oligos selector tool that can be purchase at $3885 (Three thousands, eight hundred and eighty-five dollar) and PRIMEGENS, were not tested in their report. I was able to install and configure four successfully and therefore perform the Rimour et al. (2005) experiment using them. Probesel crashed while running this experiment. The observed results were contrary to Rimour et al. (2005) findings. Array Designer 4.0 designed 1996 oligos (one oligo per gene) but only 58.31% are specific, while the 1865 oligos (one also for each CDS) returned by YODA are all specific. Goarray, the oligos selector tool of Rimour et al. (2005) expected to design more specific oligos returned 1996 oligos, but only 62.33% of these oligos are specific (oligo length is twice of 22 plus 6-mer short random linker). The summary of these results is shown in Table 1

Table 1. The results obtained from the design of oligos for 1996 CDCs of *E. cuniculi* using three tools, as shown below.

Tool	Oligo length	Total oligos	Total %	Specific probes	Specific %
Array designer 4.0	45-54	1996	100	1164	58.31
Yoda	50	1865	93.44	1865	100
Goarray	22	1996	100	1244	62.33

Table 2. The results obtained from the design of oligos for 6343 CDCs of *S. cerevisiae* using seven tools. For Goarray, oligonucleotide sequence is the concatenation of two disjoint specific short sequences (32 mer), that are complementary to their target CDNA, but separated by a very short random limker (3-6 bases).

Tool	Oligo length	Total oligos	Total %	Specific probes	Specific %
Arrayoligoselector	70	6302	99.35	5805	92.11
Goarray	32	6334	99.86	4212	66.5
Yoda	70	5809	91.58	5809	100
Oligoarray2	70	6018	94.88	3961	66.11
Array designer 4.0	65-73	6333	99.84	3493	55.16
Oligopicker	70	5892	92.89	5874	99.70
Picky	70	5385	84.90	5300	98.44

Table 3. The results obtained from the design of oligos for 1996 CDCs of *E. cuniculi* using five tools. For Goarray, oligonucleotide sequence is the concatenation of two disjoint specific short sequences (32mer), that are complimetary to their target CDNA, but separated by a very short random limker (3-6 bases).

Tool	Oligo length	Total oligos	Total %	Specific probes	Specific %
Array designer 4.0	65-73	1981	99.25	886	44.73
Yoda	70	1839	92.13	1839	100
Goarray	32	1992	99.80	950	47.69
Arrayoligoselector	70	1961	98.25	1615	82.36
Oligoarray2	70	1603	80.31	817	50.97

above. Note that for Tables 1, 2, and 3, second column indicates the oligo length and the totals number of oligos (one per gene) selected by each tool is shown in column three. Their percentages are shown in column four, while the total number of oligos that are specific for each tool is shown in column five and finally, the last column gives their percentage.

Out of the thirteen oligos selector tools accessible, eleven could be installed successfully and these include Probesel (Kaderali and Schliep, 2002), ArrayoligoSelector (Bozdech et al., 2003), OligoPicker (Wang and Seed, 2003), OligoArray 2.1 (Rouillard et al., 2003), PICKY (Chou et al., 2004), YODA (Nordberg, 2005), Array Designer 4.0 (a commercial one), Goarray (Rimour et al., 2005), OligoWiz (Niesen et al., 2003), Oliz (Chen and Sharp, 2002) and OligoSpawn (Zheng et al., 2004). Oliz could only design oligos of length 50 mer, while OligoSpawn select oligos for EST databases and its design strategy guarantee already the satisfaction of the specificity criteria as mentioned earlier on concerning tools designed using the first strategy. On several attempts, Probesel crashed on the CDS of *S. cerevisiae* and *E. cuniculi* and OligoWiz did not run properly. Nielsen confirmed that the installation performed for OligoWiz is correct but could not tell before this manuscript was submitted why OligoWiz was not running correctly on *S. cerevisiae*. The remaining seven tools was used to select oligos for 6343 genes of *S. cerevisiae* and 1996 CDS of *E. cuniculi* using the setup discussed earlier on above. OligoPicker and PICKY crashed on several attempts to find oligos for *E. cuniculi*. The oligos selected here are of length 70 and design criteria set that one have, one oligo for each CDS. Where necessary, default setting of each tool was adjusted to match another tool setting. This result is presented in Tables 2 and 3 above.

Conclusion

A comparative computational analysis (effectiveness and efficiency) has been done on tools for finding oligonucleotides needed at the front end of a high throughput microarray technology. Apart from the specificity analysis that is been done for the tools, it is observed that quite a number of the tools now design oligos for organisms at a good running time. Array Designer 3.0 (a lower version of Array Designer 4.0) spent ≅ 5040 minutes on 5864 genes of *S. cerevisiae* in the experiment performed by Nordberg (Running on a 2.0 GHz Pentium 4 processor with 512 MB RAMS), while remarkably, Array Designer 4.0 spend ≅ 16 min to find 6333 oligos for 6343 genes of *S. cerevisiae* (Running on Intel Pentium 4 with 512 MB RAM, 40 GB Hard disk and 3.0 GHz speed). Note that Array Designer 4.0 achieves 55.16% specificity (3493 oligos, one per CDS out of 6343 CDS). This is lower than 87.72% specificity achieved by Array Designer 3.0 in the Nordberg experiment. This means that specificity has been traded for time.

ACKNOWLEDGMENTS

I am grateful to Premier Biosoft for providing an evaluation license for ArrayDesigner4.0, Eric Nordberg (Virgina Bioinformatics Institute, Virgina Polytechnic Institute and State University, USA) for sending YODA and his time given for useful discussions and Lars Kaderali (Department 'Theoretical Bioinformatics', DKFZ, Heidelberg, Germany) for sending Probesel, Jean-Marie Rouillard (Department 'Chemical Engineering', University of Michigan, USA) and Henrik Bjorn Niesen (Center 'Biological Sequence Analysis' Technical University of Denmark, Denmark) for useful discussions, Hui-Hsien Chou (Department 'Computer Science' Iowa State University, USA) for granting an academic licence to use Picky, Sebastien Rimour (ISIMA, France) for sending GoArrays, Hao Chen (Department 'Pharmacology', University of Tennessee Health Science Center, USA) for sending Oliz and Stefano Lonardi (Department 'Engineering and Computer Science', University of California, Riverside, USA) for sending OligoSpawn before its availability on their website.

REFERENCES

Adebiyi EF (2007) Using suffix tree for efficient selection of unique oligos for large EST databases. Reviewed Presentation at the First Southern African Bioinformatics Workshop (SAB).

Baldi P Hatfield GW (2002). DNA microarrays and gene expression: From experiments to data analysis and modeling, Cambridge University Press.

Combes F, Lemoine S, Le Crom S (2005). Design of oligonucleotide probes for microarrays: which software for which needs?, Posters proc. JOBIM 2005.

Esiobu D (2006) Personal Communication.

Fugen Li, Gary D Stormo (2003). Selection of optimal DNA oligos for gene expression arrays, Bioinformatics, 17(11), 1067-1076.

Gerhold D, Rushmore T, Caskey CT (1999). DNA chips: promising toys have become powerful tools, Trends Biochem. Sci., 24: 168-173.

Hao Chen, Burt M Sharp (2002). Oliz, a suite of perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3' untranslated region, BMC Bioinformatics 3(27).

Henrik Bjorn Niesen, Rasmus Wernersson, Steen Knudsen (2003). Design of oligonucleotides for microarrays and perspective for design of multi-transcriptome arrays, Nuc. Acids Res. 31(13): 3491-3496.

Henrik Hornshoj, Henrike Stengaard, Frank Panitz, Christian Bendixen (2004). SEPON, a selection and Evaluation pipeline for oligoNucleotides based on ESTs with a non-target Tm algorithm for reducing cross-hybridization in microarray gene expression experiments, Bioinformatics 20(3): 428-429.

Hughes, T. R., et al. (2001) Expression profiling using microarray fabriccated by an ink-jet oligonucleotide synthesizer, Nat. Biotechnol. 19: 342-347.

Hui-Hsien Chou, An-Ping Hsia, Denise L. Mooney, Patrick S Schnable (2004). PICKY: oligo microarray design for large genomes, Bioinformatics, 20(17): 2893- 2902.

Jean-Marie Rouillard, Michael Zuker, Erdogan Gulari (2003). OligoArray 2.0:design of oligonucleotide probes for DNA microarrays using a thermodynamic approach, Nuc. Acids Res. 31(12): 3057-3062.

Kane MD, Jatkoe TA, Strumpf CR, Lu J, Thomas JD, Madore SJ (2000). Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays, Nucleic Acids Res. 28(22): 4552-4557.

Kitano H (2002). System Biology; A brief overview. Science 295: 1662-1664.

Lander ES (1999). Array of hope, Nat. Genet. 21: 3-4.

Lars Kaderali, Alexander Schliep (2002). Selecting signature oligonucleotides to identify organisms using DNA arrays, Bioinformatics 18(10): 1340-1349.

Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De la Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA (2003). Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503-1508.

Nevere NL (2001). Melting, computing the melting temperature of nuclei acid duplex, Bioinformatics 17: 1226-1227.

Nordberg EK (2005). Yoda: selecting signature oligonucleotides, Bioinformatics 21(8): 1365-1370.

Schena M (1996). Genome analysis with gene expression microarrays, BioEssays 18: 427-431.

Sebastien Rimour, David Hill, Cecile Militon, Pierre Peyret (2005). GoArrays: highly dynamic and efficient microarray probe design. Bioinformatics 21(7): 1094-1103.

Xiaowei Wang, Brian Seed (2003). Selection of oligonucleotide probes for protein coding sequences. Bioinformatics 19(7): 796-802.

Zbynek Bozdech, Jingchun Zhu, Marcin P Joachimiak, Fred E Cohen, Brian Pullian, Joseph L DeRisi (2003). Expression profiling of the schizont and trophozoite stages of plasmodium falciparum with a long-oligonucleotide microarray, Genome Biol. 4(R9).

Zheng J, Close T, Jiang T, Lonardi S (2004). Efficient selection of unique and popular oligos for large EST databases, Bioinformatics 20(13): 2101-2112.

Zuker M, Mattews DH, Turner DH (1999). Algorithms and thermodynamics for RNA secondary structure prediction, A practical guide, NATO ASI Series, Kluwer, Dordrecht.