D938–D941 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 23 October 2008*
doi:10.1093/nar/gkn761

OligoArrayDb: pangenomic oligonucleotide microarray probe sets database

Jean-Marie Rouillard* and Erdogan Gulari

Chemical Engineering Department, University of Michigan, 2300 Hayward Street, Ann Arbor, MI 48109, USA

Received September 14, 2008; Revised October 3, 2008; Accepted October 6, 2008

ABSTRACT

OligoArrayDb is a comprehensive database containing pangenomic oligonucleotide microarray probe sets designed for most of the sequenced genomes that are not covered by commercial catalog arrays. The availability of probe sequences, associated with custom microarray fabrication services offered by many companies and cores presents the unequalled possibility to perform microarray experiments on most of the sequenced organisms. OligoArrayDb contains more than 2.8 probes per gene in average for more than 600 organisms, mostly archaea and bacteria strains available from public database. On average, 98% of the annotated genes have at least one probe which is predicted to be specific to its intended target in >94% of the cases. OligoArrayDb is weekly updated as new sequenced genomes become available. Probe sequences, in addition to a comprehensive set of annotations can be downloaded from this database. OligoArrayDb is publicly accessible online at http://berry.engin.umich.edu/oligoarraydb.

INTRODUCTION

Biology's entry into the genomic era during the past decade has led to new scientific challenges concerning not only the characterization of organism's genomes, but also their expression levels. Among the various types of investigations supported by DNA microarray technologies (1–7), transcriptome analyses are the most popular (8). One can monitor the expression levels of thousands of genes in parallel, and for eukaryotes, the expression level of splicing variants (9). The availability of flexible techniques for oligonucleotide synthesis *in situ* on microarrays (10–13) and commercial custom microarray fabrication service offers the unprecedented possibility to perform gene-expression studies on organisms not represented on catalog arrays from major manufacturers. This is the case for the vast majority of all sequenced bacteria.

The ensuing challenge for researchers is to determine the oligonucleotide probe sequences to be synthesized or spotted on the microarray. We and others have developed software to design probes for transcriptome analysis (14–21) but probe design can be a cumbersome task for inexperienced researchers. Probe design is the initial step in a microarray experiment and its quality will impact the final results thus many will prefer to use predesigned probe sets rather than taking the risk of doing a suboptimal design by themselves.

Several hundreds of organisms from archaea to eukaryotes have been sequenced so far and their genomes are available on public databases. If an organism proved to be of sufficient interest to be sequenced, then there is a chance that one may want to perform gene-expression studies on that particular organism. Thus, we have designed pangenomic oligonucleotide microarray probe sets for most of the sequenced genomes especially the ones not covered by commercial catalog arrays, and compiled them into a unique database.

Here, we present OligoArrayDb, a database containing oligonucleotide microarray probe sequences that allows, in conjunction with custom oligonucleotide microarray fabrication, to envisage any transcriptome analysis on sequenced organisms. The database is freely available online from http://berry.engin.umich.edu/oligoarraydb.

MICROARRAY PROBES DESIGN

Bacterial genomic sequences were obtained from the Genbank database (22). These sequences include plasmids when available. For eukaryotes, transcript sequences from chromosomes and organelle genomes were downloaded from Genbank (22), ENSEMBL (23) or TIGR-JCVI (24).

The probe design was done using the latest version of OligoArray (20) (v3.1, Rouillard, J.-M. unpublished data). Briefly, this program searches for specific probes at the genomic scale. The probe sequence is compared to all other expressed sequences from the same organism and the thermodynamic parameters (free energy and melting temperature) are computed for all possible hybridizations between the probe and perfect or nonperfect

*To whom correspondence should be addressed. Tel: +1 734 763 4722; Fax: +1 734 763 0459; Email: jmrouill@umich.edu

complementary sequences. If all of these values fall below a predetermined threshold, the probe is considered to be specific to its target. Probes are also selected to be unable to fold into stable secondary structures that may interfere with hybridization. Any probes with low sequence complexity or long stretches of the same base are rejected.

In terms of sensitivity, and specificity, the optimal size for an oligonucleotide grown directly on a microarray and used for gene-expression analysis is comprised between 50- and 60-mers (12,25). However, the 10 nt closest to the chip's surface seem to be not involved in hybridization due to steric interference (26); therefore, there is no reason to consider these nucleotides during the design process and specially during the specificity computation, as long as the corresponding sequence or any other kind of spacers with a sufficient length will be inserted between the target sequence and the chip surface during fabrication. According to these data, we have chosen to design probes with a size comprised between 45 and 47 nt. By using a range of length, the program can fit a narrower melting temperature (T_m) range in order to achieve better hybridization uniformity. The mean GC content was computed for each input sequence and the 5% extreme values on each side were filtered out. The remaining lower and higher values were selected to set the GC content range used during design. These values were also used to determine the optimal T_m. This approach allows us to design probes with consistent thermodynamic properties for all genes. Since hybridizations are usually carried out at or below 65°C, we use this temperature as a threshold to start considering non specific hybridization, but when a genome is highly GC rich, this value is slightly increased. In some cases, gene family members are so closely related that there is no way to discriminate between pairs of them. If no specific probe is found without cross-hybridization above the threshold, then all possible nonspecific hybridizations are reported in the output. Messenger RNAs from eukaryotes are polyadenylated and since this feature is used to anchor reverse transcription during probe labeling, we have limited the search space for probes to the last 1500 nt of the input sequences for eukaryotes. This limit is to prevent picking probes in a region that would eventually not be reverse transcribed in suboptimum experimental conditions. The input sequence is searched in a 3′ to 5′ direction to give preference to probe located as closely as possible to the messenger 3′-end. For archaea and bacteria where the mRNAs are not polyadenylated, the reverse transcription is usually primed with short random primers. This will lead to a better representation of the 5′-end of the mRNAs into the cDNA population. Thus, the input sequence is searched in a 5′ to 3′ direction to preferentially pick probes close to the messengers 5′-end.

Specific probes are ranked according to their position. For prokaryotes, the probe closest to the RNA 5′-end gets the highest rank, while for eukaryotes, the probe closest to the polyA tail gets the highest rank. If no specific probe exists for a given gene, then the probe with the lowest number of nonspecific targets is ranked first.

The specificity of the probes designed for this database was assessed as follow. Briefly, probes were designed for two different genomes, yeast and a bacteria ensuring that the probe specificity was computed against both genomes. Yeast total RNAs were labeled and hybridized to a microarray containing probes for both of these genomes. After hybridization, <0.3% of the bacterial probes (18 out of 6984) showed a signal above twice the background signal. Experimental details and results are reported on the OligoArrayDb homepage.

DATABASE CONTENT

In a first run, we have attempted to design up to three probes per transcript. In order to avoid any overlap between probes, we have chosen to impose a distance between probes at least equal to half the length of the mean probe length (23 nt). This implies that for relatively short sequences, it is not possible to design more than one or two probes. At the end, we have an average number of 2.82 probes per transcript successfully processed ($n = 2\,051\,956$ transcripts with probe(s) as of 1 September 2008). We have successfully designed at least one probe for >98% of all transcripts from all organisms processed (98.3%, $n = 2\,087\,378$ transcripts from 639 organisms). More than 94% of all transcripts with probe(s) have at least one specific probe (94.7%, $n = 1\,944\,066$ transcripts with specific probes). These percentages are 84% and 72% for transcripts with 2 and 3 specific probes, respectively. In the very few cases where the design failed, it is mostly due to input sequences shorter than the probe length or to monotonous sequences containing long stretches of the same nucleotide. Overall, OligoArrayDb contains 5 778 195 microarray probes representing 2 051 956 transcripts from 639 organisms or strains as of 1 September 2008. This database is regularly updated as new sequenced genomes are released.

AVAILABILITY

OligoArrayDb is publicly available with no restriction on its usage. It can be accessed online at http://berry.engin.umich.edu/oligoarraydb. For local implementation, data flat files and a building script (postgreSQL) are also available from the home page.

RETREIVING DATA FROM OLIGOARRAYDB

The home page gives a brief description of the database purpose, including the current counts on the number of genomes, transcripts and probes. More importantly, the home page lists the available genomes, separated in three columns according to their domain of origin, archaea, bacteria and eukaryote. Within each domain, strain or species are alphabetically sorted and linked to a probe set information page.

The probe set page gives details on the input sequence source, i.e. a link to the sequence file(s) used as input as well as relevant data on the sequence composition and number of transcripts. It also gives a link to the design parameters used to generate this particular probe set. Probe set composition is described in words and visually

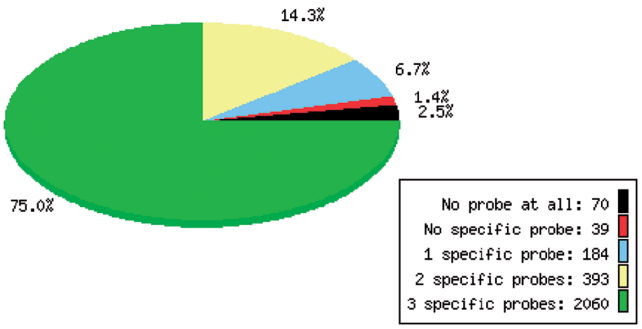

Figure 1. Number of transcripts with probes for Aeropyrum pernix K1. This genome contains 2746 genes. OligoArrayDB contains 2060 genes with three specific probes (75%, green), 393 genes with two specific probes (14.3%, yellow), 184 genes with one specific probe (6.7%, blue) and 39 genes with probes that are not specific (1.4%, red). The database does not contain any probes for 70 genes (2.5%, black).

represented as a pie chart (see Figure 1 for a typical example). Lists of genes lacking probes or specific probes are also accessible from here. Finally, this page gives a link to the probe retrieval page.

The probe page offers the possibility to choose between retrieving the full data set or a customized one. One can choose between getting all the available probes (2.8 probes per gene on average; see above) or just one or two probes per gene. In this case, output probes will be selected according to their rank (see the probe design section above), highest rank first. Then one can choose the annotations to retrieve along with the probe sequence. Possible data are gene name, function, product and locus tag if available from the input sequence file. The probe size as well as its position on the input sequence is available. Position refers to the distance between the 5′-end of the probe and either 5′- or 3′-end (the later being mostly relevant to eukaryotes) of the target. One can choose also to report predicted thermodynamic data on hybridization of the probe to its intended target (free energy of hybridization, enthalpy, entropy and melting temperature of the hybrid). Finally, one can choose to report potential cross-hybridization targets if existing. From this column, users can tell whether a probe is specific to its target or not. The probe sequence comes with different options. One can choose between getting the probe alone (45- to 47-mer) or get it flanked by up to 15 nt either on the 5′- or 3′-end of the probe. The purpose of this sequence is to be used as a spacer to increase the distance between the surface and the sequence involved in the hybridization process. The user can choose either the real sequence contiguous to the probe sequence in the input sequence or a common artificial sequence like a polyT as recommend by Guo *et al.* (27). The data will be retrieved as a <TAB> delimited file ready to import into any text editing or spreadsheet software.

DISCUSSION

We have mainly focused our database on organism of medical, agronomical and industrial interest, leaving aside for now genomes well covered by commercial arrays, including the human genome. All organisms present in OligoArrayDb have been of enough interest to be covered by a sequencing project. But due to limited market, many of them will never be covered by commercial catalog microarrays. *In situ* synthesis technologies using digital photolithography (11,13) or inkjet printing (12) provide the ultimate flexibility in microarray fabrication as these processes rely only on probe sequence files. With OligoArrayDb, we provide probes for most of the organisms for which the genome sequence is known. The availability of a full set of probe sequences, associated with *in situ* synthesis offers now the unequalled possibility to perform custom microarray experiments on these organisms. The current database contains probe sets for more than 600 organisms as of 1 September 2008 and is weekly updated as new sequenced genomes are made available.

FUNDING

National Institute of Health (1 RO1 GM06854-01A1). Funding for open access charge: College of Engineering, University of Michigan.

Conflict of interest statement. None declared.

REFERENCES

1. Ehrenreich,A. (2006) DNA microarray technology for the microbiologist: an overview. *Appl. Microbiol. Biotechnol.*, **73**, 255–273.
2. Jares,P. (2006) DNA microarray applications in functional genomics. *Ultrastruct. Pathol.*, **30**, 209–219.
3. Kato,H., Saito,K. and Kimura,T. (2005) A perspective on DNA microarray technology in food and nutritional science. *Curr. Opin. Clin. Nutr. Metab. Care*, **8**, 516–522.
4. Lettieri,T. (2006) Recent applications of DNA microarray technology to toxicology and ecotoxicology. *Environ. Health Perspect.*, **114**, 4–9.
5. Palmisano,G.L., Delfino,L., Fiore,M., Longo,A. and Ferrara,G.B. (2005) Single nucleotide polymorphisms detection based on DNA microarray technology: HLA as a model. *Autoimmun. Rev.*, **4**, 510–514.
6. Walker,M.S. and Hughes,T.A. (2008) Messenger RNA expression profiling using DNA microarray technology: diagnostic tool, scientific analysis or un-interpretable data? *Int. J. Mol. Med.*, **21**, 13–17.
7. Wiltgen,M. and Tilz,G.P. (2007) DNA microarray analysis: principles and clinical impact. *Hematology*, **12**, 271–287.
8. 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.
9. Bingham,J.L., Carrigan,P.E., Miller,L.J. and Srinivasan,S. (2008) Extent and diversity of human alternative splicing established by complementary database annotation and microarray analysis. *Omics*, **12**, 83–92.
10. Beier,M. and Hoheisel,J.D. (2005) DNA microarray preparation by light-controlled in situ synthesis. *Curr Protoc Nucleic Acid Chem*, **Chapter 12**, Unit 12.15.
11. Gao,X., LeProust,E., Zhang,H., Srivannavit,O., Gulari,E., Yu,P., Nishiguchi,C., Xiang,Q. and Zhou,X. (2001) A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res.*, **29**, 4744–4750.
12. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
13. Nuwaysir,E.F., Huang,W., Albert,T.J., Singh,J., Nuwaysir,K., Pitas,A., Richmond,T., Gorski,T., Berg,J.P., Ballin,J. *et al.* (2002)

Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.*, **12**, 1749–1755.
14. Chen,S.H., Lo,C.Z., Tsai,M.C., Hsiung,C.A. and Lin,C.Y. (2008) The unique probe selector: a comprehensive web service for probe design and oligonucleotide arrays. *BMC Bioinformatics*, **9 (Suppl 1)**, S8.
15. Feng,S. and Tillier,E.R. (2007) A fast and flexible approach to oligonucleotide probe design for genomes and gene families. *Bioinformatics*, **23**, 1195–1202.
16. Gasieniec,L., Li,C.Y., Sant,P. and Wong,P.W. (2007) Randomized probe selection algorithm for microarray design. *J. Theor. Biol.*, **248**, 512–521.
17. He,Z., Wu,L., Li,X., Fields,M.W. and Zhou,J. (2005) Empirical establishment of oligonucleotide probe design criteria. *Appl. Environ. Microbiol.*, **71**, 3753–3760.
18. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
19. Li,W. and Ying,X. (2006) Mprobe 2.0: computer-aided probe design for oligonucleotide microarray. *Appl. Bioinformatics*, **5**, 181–186.
20. Rouillard,J.M., Zuker,M. and Gulari,E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.
21. Rouillard,J.-M., Herbert,C.J. and Zuker,M. (2002) OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, **18**, 486–487.
22. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
23. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
24. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res.*, **29**, 123–125.
25. Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
26. Shchepinov,M.S., Case-Green,S.C. and Southern,E.M. (1997) Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res.*, **25**, 1155–1161.
27. Guo,Z., Guilfoyle,R.A., Thiel,A.J., Wang,R. and Smith,L.M. (1994) Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res.*, **22**, 5456–5465.