



OligoArray: genome-scale oligonucleotide design for microarrays

Jean-Marie Rouillard¹, Christopher J. Herbert² and Michael Zuker³

¹Department of Chemical Engineering, University of Michigan, Ann Arbor, MI 48109-2136, USA, ²Centre de Génétique Moléculaire du CNRS, Av de la Terrasse, 91198 Gif-sur-Yvette, France and ³Department of Mathematical Sciences, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA

Received on June 22, 2001; revised on October 9, 2001; accepted on October 11, 2001

ABSTRACT

Summary: OligoArray is a program that computes gene specific and secondary structure free oligonucleotides for genome-scale oligonucleotide microarray construction or other applications.

Availability: The program code is distributed under the GNU General Public License and is freely available for non-profit use via request from the authors.

Contact: jean-marie.rouillard@cgm.cnrs-gif.fr

Supplementary information: <http://berry.engin.umich.edu/oligoarray>

DNA microarrays enable the expression of thousands of genes to be monitored in parallel. For the production of microarrays, DNA can either be synthesized on a solid support (Lipshutz *et al.*, 1999) or deposited in a pre-synthesized form onto a suitable surface. In this case the DNA can be in the form of PCR products (Skena *et al.*, 1995) or oligonucleotides (Hughes *et al.*, 2001). These authors and others (Kane *et al.*, 2000) have demonstrated that oligonucleotide microarrays compare well with cDNA microarrays and that a single oligonucleotide per gene is sufficient to monitor gene expression.

The emergence of new flexible oligonucleotide technologies in microarray fabrication that only require sequence data (Hughes *et al.*, 2001) and the availability of an increasing number of sequenced genomes has prompted us to develop a program to design oligonucleotides for microarrays. Here we present OligoArray, a program that computes gene specific and secondary structure free oligonucleotides for genome-scale oligonucleotide microarray construction.

OligoArray is written in Java 1.2 and runs under all platforms supporting Java and the Blast program. The user, via a graphical interface (see **Supplementary information**), can configure parameters such as length, number per sequence, maximal distance from the end

of the sequence, melting temperature range, threshold to reject secondary structures, sequence tags to add to the 5' and/or 3' ends of the oligonucleotide and prohibited sequences. Prior to running OligoArray, all sequences that need to be processed for an organism are saved in a file with a FastA format. These sequences can be mRNA sequences, CDS, or exon sequences, depending on which part of the sequence the search will be restricted to. For each entry in the input file, the sequence is read backwards from the 3' end using a moving window length equal to the length of the oligonucleotide. This window sequence is first examined for the presence of prohibited sequences. This allows the user to avoid the presence of sequences that may be needed for other applications. Then, the oligonucleotide's specificity is tested against a database containing all transcribed sequences from the same organism. This database needs to be built before OligoArray is run and should be non-redundant (see **Supplementary information**). The Blast program (Altschul *et al.*, 1997), available from the NCBI, is run using the -F F and -S 1 options to consider low-complexity sequences and only the transcribed strand of each sequence respectively. The threshold of specificity necessary for an oligonucleotide to be accepted was designed to be more stringent than that recommended by Kane *et al.* (2000) and is a function of the level of identity. For similarity spanning more than 50 Nucleotides (nt), the identity has to be less than 50%. For similarity spanning 36–50 nt and 15–35 nt, the thresholds are 60 and 70% respectively. Perfect identities smaller than 15 nt are accepted, and in a second round, a perfect identity of 15 nt is tolerated. If no oligonucleotide fulfills these criteria, the sequence is considered to belong to a family and will be processed differently as described below. Sequences that pass the specificity test are examined for the absence of strong secondary structure that could interfere with hybridization. The melting temperature of all possible

secondary structures is computed by the Mfold server (Zuker *et al.*, 1999) using thermodynamic parameters from SantaLucia (1998), a sodium concentration of 1M and a temperature of 50°C to compute structure free energy. The oligonucleotide is accepted if no structure presents a melting temperature above the threshold defined by the user. If the current oligonucleotide sequence does not fulfill these criteria, the sequence window is moved iteratively by 10 nt to the 5' end of the sequence until a suitable oligonucleotide or the maximal authorized distance separating the 5' end of the oligonucleotide to the end from the sequence is reached.

If no oligonucleotide is found for a sequence because of a high level of identity with other sequences, this usually means that the sequence is part of a gene family. We define a family as a group of sequences that share more than 90% identity along at least the length of the oligonucleotide minus 5 nt. The percentage of similarity and the number of similar sequences may vary along the length of the current sequence. An oligonucleotide chosen at 500 nt from the 3' end can be similar to fewer sequences than an oligonucleotide chosen 100 nt from the 3' end and thus will belong to a smaller family. The program will search for an oligonucleotide in the region corresponding to the smallest family of sequences. Other sequences excluded from the family as we defined it, but possibly involved in the overall hybridization to this oligonucleotide, are reported in the output file.

Once the input file has been selected, OligoArray does not need further action from the user. Sequences are automatically transferred to Blast and Mfold programs and the results are saved in two output files. The main one is in a FastA format and contains the oligonucleotide sequences. The second file contains rejected sequences. Examples are provided in the **Supplementary information**.

To test the program, we used it to select oligonucleotides from the genome of the yeast *Saccharomyces cerevisiae* (6343 genes). In a first round, oligonucleotides representing 6334 genes were successfully designed (length = 50, T_m range = $87 \pm 5^\circ$). With these criteria, oligonucleotides could not be selected for nine genes. For eight of them, oligonucleotides could be designed using a lower T_m range ($80 \pm 5^\circ$) since they are AT rich. No specific oligonucleotide could be found for the last gene due to its short length (YOR008w-B, 102 nt). The mean distance between the oligonucleotide and the 3' end of the CDS is 237 nt and this distance is smaller than 500 nt for 90% of the

genes. Among the 6342 oligonucleotides obtained, 5883 represent a unique gene, 65 represent a unique gene but can present some cross hybridization with other mRNAs and 394 represent a family of more than one gene as described above. These results are consistent with gene duplication in yeast (Coissac *et al.*, 1997). This computation was done in less than 1 day on a 700 MHz processor.

Specificity selection is based on the percentage of similarity inside a sequence window. Furthermore thresholds to reject non specific oligos are more stringent than recommended by Kane *et al.* (2000). This probably leads the program to reject a few oligonucleotides that should not be rejected. In order to reduce this limitation, future improvements will concern the implementation of an algorithm to compute specificity based on thermodynamic parameters of all possible hybridizations between oligonucleotides and expressed sequences from the same organism. Furthermore, a version using variable oligonucleotide length is currently under development.

REFERENCES

- 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 programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Coissac,E., Maillier,E. and Netter,P. (1997) A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.*, **14**, 1062–1074.
- 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. *Nature Biotechnol.*, **19**, 342–347.
- 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 (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
- Lipshutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet.*, **21**, 20–24.
- SantaLucia,J. Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
- 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.
- Zuker,M., Mathews,D.H. and Turner,D.H. (1999) *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide*, NATO ASI Series, Kluwer, Dordrecht.