Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W611–W615***
doi:10.1093/nar/gki399

OligoWiz 2.0—integrating sequence feature annotation into the design of microarray probes

Rasmus Wernersson* and Henrik Bjørn Nielsen

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800, Lyngby, Denmark

Received February 14, 2005; Revised and Accepted March 14, 2005

ABSTRACT

OligoWiz 2.0 is a powerful tool for microarray probe design that allows for integration of sequence annotation, such as exon/intron structure, untranslated regions (UTRs), transcription start site, etc. In addition to probe selection according to a series of probe quality parameters, cross-hybridization, T_m, position in transcript, probe folding and low-complexity, the program facilitates automatic placement of probes relative to the sequence annotation. The program also supports automatic placement of multiple probes per transcript. Together these facilities make advanced probe design feasible for scientists inexperienced in computerized information management. Furthermore, we show that probes designed using OligoWiz 2.0 give rise to consistent hybridization results (http://www.cbs.dtu.dk/services/OligoWiz2).

INTRODUCTION

The appearance of next generation micro-array technologies, with emphasis on high-density, low cost custom oligonucleotide-arrays, such as the NimbleExpress (Affymetrix, CA), together with the increasing number of sequenced genomes, opens up a new world of opportunities for the biologist. Using customized arrays it now becomes feasible to do different types of experiments, e.g. expression analysis of exciting newly sequenced organisms, special purpose studies, such as alternative splicing (1,2), mapping of untranslated regions (UTRs) and screening intergenic regions for novel transcripts.

In order to fully exploit the potential of these advances, it is crucial to have access to probe design tools that provide the required flexibility to design probes for this wide range of purposes. Such a tool should also provide a good overview of the different aspects of probe design, e.g. probe quality parameters, the placement along the target transcripts and must also aid in identifying high quality probes.

OligoWiz 1.0 has since its release two years ago (3) showed its strength as a very flexible probe design tool. The scoring scheme for probes, the flexible weighting system and the availability of a range of genome databases, have made OligoWiz 1.0 popular for the design of custom oligonucleotide-arrays. Currently, ~50 000 genes are submitted to the OligoWiz 1.0 server every month.

However, OligoWiz 1.0 is primarily build for selecting one single long probe (50–70 bp) per gene, aimed at traditional gene expression analysis. The valuable feedback we have received from the users of OligoWiz 1.0, as well as our own experience suggested that there was a demand for an expansion that could automate the selection of multiple probes per transcript. Also, the effort of designing special purpose microarrays is in our experience a tedious and demanding task. Therefore, we have integrated the ability to work with sequence feature annotation directly into OligoWiz 2.0, as part of the scheme for automatic placement of multiple probes. The advanced rule-based selection of probes is one of the most important new features in OligoWiz 2.0.

A small number of general-purpose programs for microarray oligonucleotide probe selection have been published (4–6). These programs, much like OligoWiz 1.0 (3), feature some kind of quality assessment of the probes available for detecting a transcript. Typically through detection of possible cross-hybridization and some physical/chemical properties of the probes, like melting temperature of the probe:target bond. In addition, some programs estimate the folding potential of the probes using mfold (5–7).

Furthermore, a number of special purpose probe designs have been reported without providing a general method (7–9). Only one of the available programs describe standard protocol for placing multiple probes (6) within each transcript and no program is available for placing probes relative to sequence annotation, such as exon/intron structure, UTRs, transcription start site, etc.

Since one of the goals of OligoWiz is to encourage the user to experiment with the array design, it has always been the aim of OligoWiz 2.0 to deliver the result in a reasonably short time scale—for example, the processing of the ~5600 transcripts

*To whom correspondence should be addressed. Tel: +45 45252489; Fax: +45 45931585; Email: raz@cbs.dtu.dk

in the Yeast genome takes ~45–60 min, depending on the server load.

QUICK REVIEW: SCORING SCHEME

OligoWiz 2.0 utilizes a set of scores each describing how well suited each possible probe, along the transcript sequences is for use as a DNA microarray probe, according to the following criteria: Cross-hyb, Delta-Tm, Low-complexity, Position and Folding. Each score has a value between 0.0 (not suited) and 1.0 (well suited). All of these scores are combined using a weighting scheme to form a Total score for each possible probe. The Total score is used for selecting the best-suited probe(s).

A comprehensive description of the algorithms used for calculating the scores—Cross-hybridization (previously 'Homology'), ΔT_m and Low-complexity is found in the OligoWiz 1.0 paper (3).

The following changes in the probe scoring scheme have been introduced since OligoWiz 1.0:

(i) The 'GATC-only' score has been removed; since its filtering behavior can be mimicked through the rule-based selection of probes.
(ii) An effort has been done to parameterize the position score for both prokaryotes and eukaryotes. The score now supports five modes: Poly-A priming, random priming, linear 5′ or 3′ preference and linear mid preference.
(iii) A completely new folding score has been implemented and is described in detail below.

FOLDING SCORE

To estimate to what extent the probes are available for hybridization with the target, the self-annealing ability must be estimated. For probes that are attached to the array support in one end, this is equal to a probe folding prediction. The main reason for not including a score for folding (self-annealing) in OligoWiz 1.0 was the overwhelming computational burden of secondary structure calculation using programs, such as mfold (10–12). Therefore an alternative and faster algorithm to estimate the folding energy, utilizing the overlapping nature of consecutive probes along a transcript, was developed.

Initially the full transcript sequence is translated into a 16-letter alphabet representing the dinucleotides. Then a super-alignment matrix covering the whole transcript is built, using stacking energies for the dinucleotides as substitution scores (Figure 1). To gain speed a hash entry for each of the 16 possibly dinucleotides, containing an alignment row in the super-alignment matrix, were used to fill the respective rows of the super-alignment matrix.

Subsets of the super-alignment matrix were then used to calculate the folding of the consecutive probes along the length of the transcript, by dynamic programming (local alignment) (Figure 1, triangles). The dynamic programming algorithm is allowed to utilize the path graph of the previously calculated probe, which limits the required calculations to include a few new rows in the new path graph and thereby speeds up the calculation significantly.

The algorithm allows gaps and summation of multiple independent folds with folding energy less than −4 kcal/mol to return an overall folding energy.

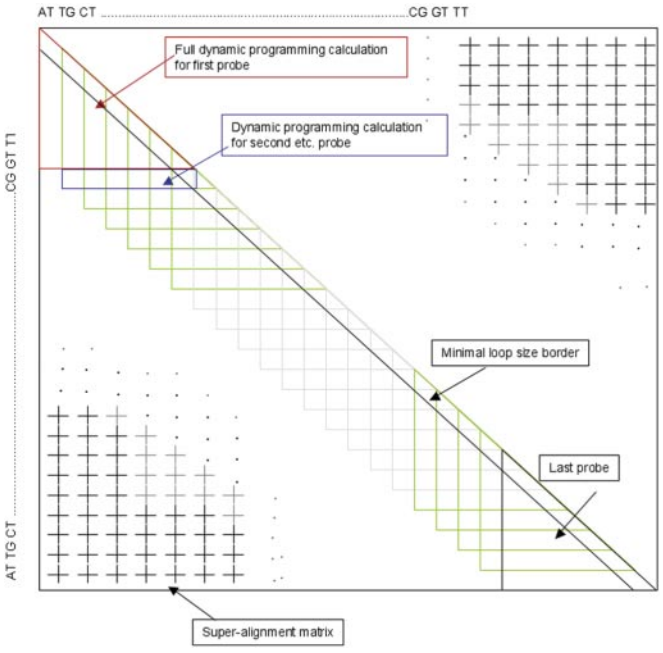

Figure 1. Diagram of the folding prediction algorithm in OligoWiz 2.0.

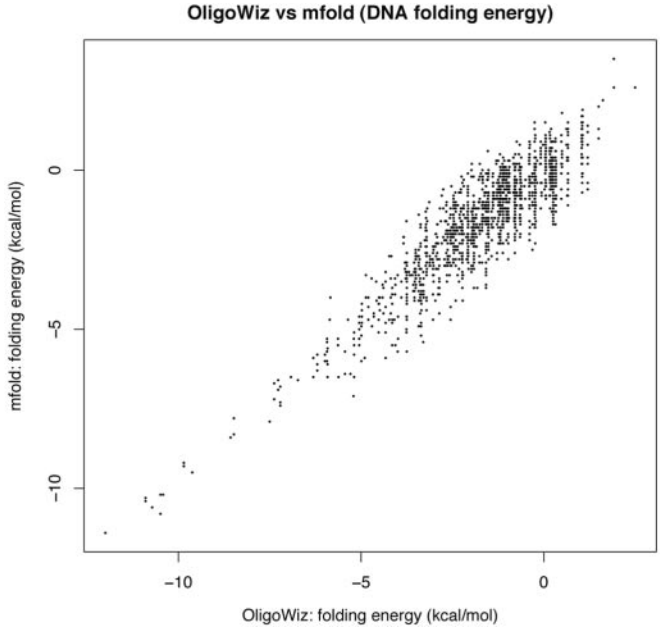

Figure 2. Scatterplot showing the folding energy as estimated by OligoWiz 2.0 versus the estimate from mfold. Folding energy of all 25 bp subsequences of three *S.cerevisiae* transcripts (acc: YOR084W, YDL144C and YFR018C, of 1071 bp, 1164 bp and 1092 bp, respectively) were estimated. Especially for the strongly folded probes, the correlation is high.

This simple algorithm estimates folding energies for subsequences (potential probes) along an input sequence 500–1000 times faster than nafold [the core program of mfold (11)], resulting in a time consumption of ~1.5 s for all 25 bp subsequences of a 1000 bp input sequence, when run on the OligoWiz 2.0 server.

To evaluate the precision of the folding algorithm, the estimated folding energy for all 25 bp subsequences of three *Saccharomyces cerevisiae* transcripts (acc: YOR084W, YDL144C and YFR018C, of 1071, 1164 and 1092 bp respectively) was compared with mfold estimations. The two folding energy estimates are plotted against each other in Figure 2. The two estimates have an overall Pearson correlation of 0.89 and for subsequences, estimated to have a folding energy lower than −6 kcal/mol, the correlation is 0.986.

For the OligoWiz 2.0 server the folding energy is converted into a 'Fold score' that ranges from 1 to 0, where 1 is 'no significant folding', and 0 is 'strong folding'.

$$\text{Foldscore} = 1 - \left(\frac{F}{-k} + \frac{L}{k^2} \right),$$

where F is the free energy of the folding (kcal/mol), L is the probe length and k is a constant (default 20).

INTEGRATING SEQUENCE FEATURES INTO THE PROBE DESIGN

In order to place probes relative to sequence features, such as intron/exon structure, the user has the option of supplying a sequence feature annotation string along with each input sequence.

The annotation string consists of a single-letter annotation code, one letter for each position in the input sequences. As an example we use the letter 'E' to annotate nucleotides which are part of an exon and the letter 'I' for those, which are part of an intron in some of the example datasets available at the OligoWiz 2.0 website, http://www.cbs.dtu.dk/services/OligoWiz2.

A combined sequence and annotation file can easily be custom made or extracted from GenBank files using the FeatureExtract server (15) (URL: http://www.cbs.dtu.dk/services/FeatureExtract), which was build for this purpose. The file format is described in detail at the OligoWiz 2.0 website.

Figure 3. Visualizing sequence feature annotation. 1: Graphs visualizing the suitability scores for each potential probe along the transcript. 2: Bar representing the entire transcript. The default color code will show exons in green and introns in blue. 3: Detailed probe information—DNA and annotation string. Key to the annotation string: 'D': donor site, 'I': intron, 'A': acceptor site, '(': start of exon, 'E': exon, ')': end of exon.

The sequence feature annotation is visualized in the graphical interface (Figure 3). Combined with the rule-based placement of probe described in the next section, this enables the scientist to apply different placement strategies and immediately inspect the placement in the context of the graphical representation.

RULE-BASED PLACEMENT OF PROBES

We have implemented a rule-based method of probe placement that builds upon the existing scheme of parameter scores. The rationale behind this approach is to make it possible (i) to place multiple probes within each transcript according to the desired distance criteria and (ii) to take sequence annotation into account (Figure 3).

For each sequence, the steps in the probe placement algorithm are as follows:

(i) If any filters have been defined, mask out probe positions that do not fulfill the criteria (for details, see below).
(ii) Place a probe at the position with the highest Total score.
(iii) Mask out surrounding positions, as defined by the minimum probe distance setting.
(iv) If the maximum number of probes per sequence has not been reached, go to step 2.

The search can be restricted to sub-sets of the input sequence of interest, by defining a set of conditions that must be present and/or absent in the sequence feature annotation or the transcript sequence itself. These conditions are defined with regular expressions (advanced text-based matching), which are used to create a filter that defines the sub-set of the transcript that will be considered during the iterative probe placement. Detailed instructions on how to take advantage of sequence feature annotation in combination with rule-based placement of probes can be found on the OligoWiz 2.0 website.

CONSISTENT HYBRIDIZATION

Evaluating a probe design is not an easy task and furthermore it is often considered too costly. Here, we decided to evaluate the OligoWiz 2.0 probe design by designing 7–8 probes of 24–26 bp for each of 3278 *Aspergillus nidulans* genes. 3278 correspond to the most well annotated genes of *A.nidulans* (annotated by the Broad Institute). A microarray containing these probes was synthesized *in situ* on a geniom one microarray system [Febit, Manheim, Germany (13)]. Labeled aRNA from 12 independently grown *A.nidulans* samples were hybridized onto the array. The Pearson correlation between the probe intensity measures and the gene expression index (14) through the 12 samples were used as a measure of probe consistency. The average probe correlation to the expression profile was 0.85 (Figure 4). This correlation showed clear intensity dependence, with high correlation for significantly expressed genes and less correlation for genes expressed close to the background level. A very conservative interpretation of these results is that the probes designed by OligoWiz 2.0 give internally consistent signals.

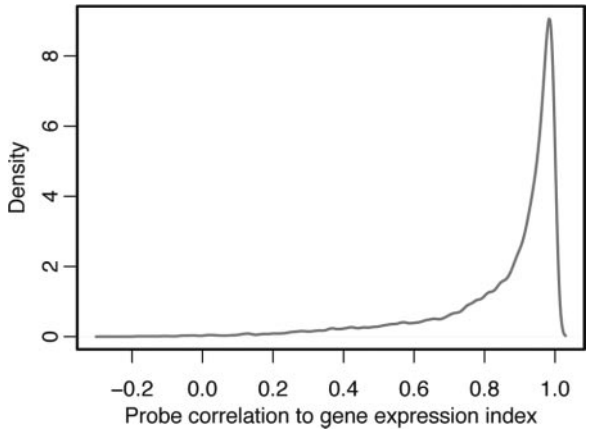

Figure 4. The distribution of probe correlations to the gene expression index (14), through 12 independent measurements of 3278 genes using 7–8 probes per gene. The plot illustrates that the majority of the probes agree with the gene expression index.

ACKNOWLEDGEMENTS

A grant from The Danish Technical Research Council (STVF) for the 'Systemic Transcriptomics in Biotechnology' financed this work. We thank Hanne Jarmer for suggesting numerous improvements for the program. Funding to pay the Open Access publication charges for this article was provided by a grant from The Danish Technical Research Council (STVF) for the 'Systemic Transcriptomics in Biotechnology'.

Conflict of interest statement. None declared.

REFERENCES

1. Clark,T.A., Sugnet,C.W. and Ares,M.,Jr (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.
2. Wang,H., Hubbell,E., Hu,J., Mei,G., Cline,M., Lu,G., Clark,T., Siani-Rose,M.A., Ares,M., Kulp,D.C. *et al*. (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, **19**, i315–i322.
3. Nielsen,H.B., Wernersson,R. and Knudsen,S. (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res*., **31**, 3491–3496.
4. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
5. 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.
6. Reymond,N., Charles,H., Duret,L., Calevro,F., Beslon,G. and Fayard,J.M. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20**, 271–273.
7. Rimour,S., Hill,D., Militon,C. and Peyret,P. (2004) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics*, **21**, 1094–1103.
8. Mrowka,R., Schuchhardt,J. and Gille,C. (2002) Oligodb—interactive design of oligo DNA for transcription profiling of human genes. *Bioinformatics*, **18**, 1686–1687.
9. Emrich,S.J., Lowe,M. and Delcher,A.L. (2003) PROBEmer: a web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Res*., **31**, 3746–3750.
10. Zuker,M. (1994) Prediction of RNA secondary structure by energy minimization. In Griffin,A.M. and Griffin,H.G. (eds), *Computer Analysis of Sequence Data*. Humana Press, Inc., Totowa, NJ, Vol. 25, Part II, pp. 267–294.
11. Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In Barciszewski,J. and Clark,B.F.C. (eds), *RNA*

Biochemistry and Biotechnology. NATO ASI Series, Kluwer Academic Publishers.
12. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
13. Baum,M., Bielau,S., Rittner,N., Schmid,K., Eggelbusch,K., Dahms,M., Schlauersbach,A., Tahedl,H., Beier,M., Güimil,R. *et al.* (2003) Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res.*, **31**, e151.
14. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
15. Wernersson,R. (2005) FeatureExtract—extraction of sequence annotation made easy. *Nucleic Acids Res.*, **33**, W567–W569.