Published online 21 February 2010

Nucleic Acids Research, 2010, Vol. 38, No. 10 ***e117***
doi:10.1093/nar/gkq110

Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments

Laurent Jourdren[1], Aurélie Duclos[1], Christian Brion[1,2], Thomas Portnoy[1,2], Hugues Mathis[2], Antoine Margeot[2] and Stéphane Le Crom[1,*]

[1]Institut de Biologie de l'École normale supérieure (IBENS), Institut National de la Santé et de la Recherche Médicale U1024, Centre National de la Recherche Scientifique UMR8197, 75005 Paris and [2]IFP, Département Biotechnologie, 1-4 Avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France

Received October 14, 2009; Accepted February 8, 2010

ABSTRACT

Despite the development of new high-throughput sequencing techniques, microarrays are still attractive tools to study small genome organisms, thanks to sample multiplexing and high-feature densities. However, the oligonucleotide design remains a delicate step for most users. A vast array of software is available to deal with this problem, but each program is developed with its own strategy, which makes the choice of the best solution difficult. Here we describe Teolenn, a universal probe design workflow developed with a flexible and customizable module organization allowing fixed or variable length oligonucleotide generation. In addition, our software is able to supply quality scores for each of the designed probes. In order to assess the relevance of these scores, we performed a real hybridization using a tiling array designed against the *Trichoderma reesei* fungus genome. We show that our scoring pipeline correlates with signal quality for 97.2% of all the designed probes, allowing for a posteriori comparisons between quality scores and signal intensities. This result is useful in discarding any bad scoring probes during the design step in order to get high-quality microarrays. Teolenn is available at http://transcriptome.ens.fr/teolenn/.

INTRODUCTION

The development of high-throughput sequencing technologies challenges the classical application fields conquered by microarrays during the last decade (1). To face this new competition, microarray providers concentrate their efforts on the improvement in feature density and sample multiplexing. Thus, biochips became flexible enough to face a range of applications not yet easily available with high-throughput sequencing. These applications are clearly complementary to the ones found with sequencing such as genotyping (2). Microarrays are still attractive to perform comparative genomic hybridization (CGH) or high-resolution transcriptome analyses on small microorganism genomes. Affordable multiplexing arrays are also of great interest for system biology approaches where kinetic experiments are needed to create dynamic models.

However, the oligonucleotide design step is a real bottleneck for all these applications. Despite all the existing probe-design software, this step remains complex and is considered something of a black box for many microarray users. Designing probes for microarrays requires dealing with several parameters, such as sensitivity and specificity. Sensitivity is defined by the strength with which a probe binds to its target sequence. It influences the level of the signal read from the microarray and the relevance of the obtained information. Specificity is defined according to the ability of the probe to bind to non-target sequences in the hybridization sample. Cross-hybridization is usually one major source of non-specificity. The selection of the best oligonucleotide design solution is therefore even more difficult (3) since it is often dependent on the expected application. Several solutions that allow the design of tiling arrays with long oligonucleotides are freely available for academics. First, OligoTiler (4) and Lipson *et al.* (5) algorithms optimize the tiling path (succession of oligonucleotides along the genome) in order to obtain the most even distribution of probes. OligoTiler is able to find probes even in repeated regions but offers a limited control on the sensitivity and the specificity of

*To whom correspondence should be addressed. Tel: +33 1 44 32 23 72; Fax: +33 1 44 32 39 88; Email: lecrom@biologie.ens.fr

the designed oligonucleotides. The Lipson *et al.* solution works on a subset of oligonucleotides with an acceptable quality, and then selects probes that ensure the most evenly distributed tiling path. For both these solutions, probe selection is therefore mainly based on probe position and not on their quality. In contrast, Tileomatic (6) and ArrayDesign (7) focus mainly on probe quality; either by selecting the probe of best quality in each window of the tiling path (ArrayDesign), or by using an implementation of the shortest path algorithm (Tileomatic). The specificity calculation, which is necessary to estimate probe quality, is done using either a suffix array approach (Tileomatic) or a uniqueness score calculation based on minimum unique prefix count for each oligonucleotide (ArrayDesign). Both these programs allow the design of oligonucleotides in repeated regions if a large unique overlapping probe can be found. Among these four solutions, only OligoTiler and ArrayDesign are implemented. Each of these software programs runs according to its own properties. Both lack flexibility and are not able to assign probe quality value to each designed probe.

Here, we describe a new flexible and universal probe design solution (Teolenn) based on an open module system approach. This software answers everyone's needs and remains user-friendly. In addition, our design program assigns a quality score to each designed probe so that a posteriori filtering steps and correlations can be made. We compared the results of several tests we performed with the ones obtained from other available solutions. Finally, we designed a microarray against a fungal genome in order to perform hybridizations; this allowed us to compare the quality scores we calculated with the intensity of the signal obtained from a genomic DNA-hybridization experiment.

MATERIALS AND METHODS

Teolenn probe parameter calculation

After the creation of the library of all possible probes, we used the SOAP software (8) to detect redundant oligonucleotides. Teolenn launched SOAP v1 with a seed size of 12, and a limit of five maximum mismatches on a read. Only probes with a unique match were conserved in the library. Complexity is evaluated using the masked genome by counting the number of masked bases for each probe. T_m values are calculated using the nearest neighbour thermodynamic model (9). The 'uniquesub' function of genome tools (7) is used to calculate the uniqueness of each probe in the library.

Trichoderma reesei probe design

We downloaded the unmasked fasta file of the *T. reesei* genome v.2.0 from the Department of Energy Joint Genome Institute website (JGI): http://genome.jgi-psf.org/Trire2/Trire2.home.html. We designed a *T. reesei* tiling array with 60 mer oligonucleotides (oligo length) each 150 bp (oligo distance) using OligoTiler from its web interface (http://tiling.gersteinlab.org/OligoTiler/oligotiler.cgi). The advanced parameters were set up as follows: 'IR region' = 5, 'IR require' = 3, 'repeat region overlap' = 4.

ArrayDesign software was retrieved from the author website (http://www.ebi.ac.uk/~graef/arraydesign/). We created sequence windows of 150 bp every 149 bp along the 87 scaffolds using Exonerate tools (10). Minimal unique prefix was computed with the MAX_PREFIX_LENGTH variable set at 15. Finally, we launched the oligonucleotide selection using the following parameters: minimal uniqueness score = 0, offset to shift window over unit for uniqueness score = 1, T_m value range = 60–80°C, G number cut-off = 15, percent palindromic filter = 40%, maximum number of synthesis cycles allowed = 185. No deviation of probe length was allowed.

The *T. reesei* design was done with Teloenn using a MAX_PREFIX_LENGTH set to 15 for the uniqueness calculation made by genome tools. No filters were applied after probe parameter calculations. In order to obtain the probe quality score, the calculated parameters were weighted as follows: 0.4 for T_m, 0.3 for uniqueness, 0.2 for GC content and 0.1 for complexity. To get final oligonucleotide scores, a weighting of 0.75 was assigned to quality scores, and a weighting of 0.25 was assigned to position scores. We also performed a variable probe length design using the same parameters than above and allowing a four-base variation of the total probe length.

Probe design comparison

We used the Unafold suite (11) to compute the free energies of the most probable secondary structures for each oligonucleotide. We used the melt.pl script with a hybridization temperature fixed to 65°C. We set DNA concentration to 0.00001 M, sodium to 1 M and magnesium to 0 M. The interval between oligonucleotides was calculated by measuring the distance between the first position (start) of two successive oligonucleotides. To calculate the number of designed probes per transcript, we first retrieved the 'Filtered Models' transcript file from the JGI website. We selected all 'exon' features from this file, and we calculated the number of oligonucleotides fully included in each of these annotated exons. For each transcript, we merged all exon information, and retrieved the total number of oligonucleotides.

To estimate Kane's parameters, we launched WU-BLAST (12) on each oligonucleotide with the *T. ressei* reference genome using the following parameters: expectation threshold for reporting database hits = 1.2 (E), seed word length for the ungapped BLAST algorithm = 11 (W), negative penalty score for mismatch nucleotides in the BLASTN search mode = –3 (N), penalty for a gap of length one = 3, per-residue penalty for extending a gap = 3. Gapped alignments were not allowed to be created. From the output files, we counted the number of hits where the percentage of identity exceeded 75% with an alignment length of 60 bp.

Comparative genomic hybridization

The microarray data and the related protocols are available at the GEO web site (http://www.ncbi.nlm.nih

.gov/geo/) under accession number: GSE17752. Briefly, chromosomal DNA from *T. reesei* QM6a strain was prepared as described previously (13). Two times 4 µg of genomic DNA were labelled with Alexa 555 or Alexa 647 using the BioPrime Total aCGH-labelling kit (Invitrogen). The two samples were then mixed together and hybridized according to the oligo aCGH/ChIP-on-Chip hybridization kit (Agilent) on a 244k array ordered from Agilent. The array was read using a GenePix 4000B scanner (Molecular Devices) and signal analysis was done using the GenePix Pro 6.1 software. Data pretreatment was applied on each result file to discard GenePix flag and saturating spots. The data were normalized without background subtraction by the global Lowess method performed with the Goulphar software (14).

The GenePix Pro analysis software flagged 'not found' spots when the 'align blocks' algorithm was not able to locate features on the slide. A spot was labelled as 'detectable' when the raw mean intensities were above the background. The background threshold was calculated by adding 2 SDs to the average intensity of all the 'not found' features. The detection threshold was 6.06 in our CGH experiment.

RESULTS

Software implementation

Teolenn works as a four-step workflow (Figure 1). Each step uses the modular organization we set up and is fully customizable. New parameter calculations or new filters can be added to fit the final user needs.

Step 1—Probe library creation. The first step of the workflow consists of creating a library of all the possible probes along the genome sequence of the organism of interest against which the probes have to be designed (called 'reference'). These probes can be designed either with a fixed or variable length. An analysis of probe redundancy is performed after the library creation, in order to discard all the probes that are found several times in the collection. The analysis is set to search for similarity among probes, allowing several mismatch positions.

Step 2—Probe parameter calculation. For all the probes of the created library, the software calculates the parameters needed to estimate oligonucleotide quality. These calculations are implemented as modules. Any developer can add new parameters to the workflow by programming its own module. In its current version, the workflow can calculate the melting temperature (T_m), the GC percent, the complexity and the uniqueness for each probe. T_m is calculated with the nearest neighbour method (9); complexity is measured using the masked reference genome; and uniqueness is estimated based on GenomeTools (7).

Step 3—Probe filtering. This step allows customization of the workflow in order to obtain the final oligoset that best fits the user's needs. This filtering process is also based on modules and can therefore be conducted in several ways.

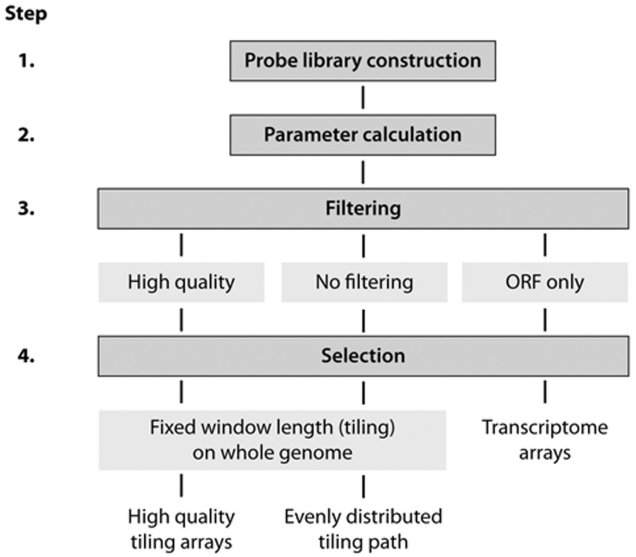

Figure 1. Probe design workflow. The probe design workflow is composed of four steps (see text for details). Here we show several application possibilities depending on the parameters used for the filtering and selection steps. For example, to design a classical transcriptome microarray, the filtering step will keep only the probes from the library that are found in ORFs from genome annotations. Next, the selection step will select the best probes (or several) for each ORF according to user's specifications. At the opposite, one can make a design without any a priori on genome annotation for tiling arrays. On this figure, we show two different ways to design tiling arrays. The first one creates 'high-quality' tiling arrays by filtering low quality probes from the library out (e.g. high or low GC content, non-homogenous melting temperature, etc.). In each window with a fixed length, the best probe will be selected only among the highest quality oligonucleotides of the library. With such a design it is possible that several windows do not get to any corresponding probe. At the opposite, if one wants to favour an even distribution of the tiling path, the constraint on probe filtering can be relaxed in order to keep most of the probes from the library. All the windows will then have a probe selected, though with lower quality parameters.

New filters that give the design workflow new abilities can be set up by developers. For example, the user can apply filtering ranges on the calculated parameters in order to keep only high-quality probes. Conversely, a user who wants to design tiling arrays with a homogenous distribution of probes can keep all the possible probes by using less stringent probe quality filters.

Another possibility may be to set filters that are based on genome annotations in order to select probes that only cover ORFs, small RNAs and so on.

Step 4—Probe selection. The final step of the probe design workflow is dedicated to the selection of the best probe in each window. The idea of a genomic window is very flexible here, as a window can be set up to a fixed or a variable length. A score position is calculated for each oligonucleotide of the selected windows in a way that the most central probe will get the best score.

A final oligonucleotide score is then calculated by combining this position score and the quality score calculated during Step 2. The best probe for each window for microarray design is the one that gets the highest final score. All these score calculations can be

balanced to modify the relative weight of each parameter. Teolenn can output the results in various file formats (plain text, fasta files, GFF or custom output) to provide ready-to-use data. For example, a GFF output in a genome browser like gBrowse (15) exhibits designed probe positions on the genome.

Probe design tests

In order to test our probe selection workflow, we designed a tiling microarray against the genome of *T. reesei (Hypocrea jecorina)* (16). Thanks to its ability to degrade plant cell wall polysaccharides efficiently by synthesizing and secreting cellulase enzymes, this fungus is increasingly being investigated in various fields of biotechnology, especially in biofuel production from lignocellulosic biomass. However, it is not available from microarray providers, which makes it a good model for organisms that require the use of probe design software.

We chose to cover the whole genome of *T. reesei* (34 Mb) with a 244 000-feature microarray. This led us to design one oligonuceotide (60 mer each) every 150 bases. The library construction step resulted in 0.35% probes discarded as 33 334 323 probes were produced out of the 33 449 658 possible oligonucleotides. These discarded probes were actually found to be redundant within the genome of *T. reesei* and consequently not suitable for microarrays. We next calculated the parameters needed for the estimation of probe sensitivity and specificity. Figure 2 shows that the distribution of probe parameters was very homogenous within the library. Most of the probes (75.64%) that were found in the library had a melting temperature of between 70 and 80°C; 89.78% of them had a GC percent of between 40 and 70%. This analysis of probe parameter distribution helped to set the thresholds that filter out bad probes from the library.

As we decided to favour a homogenous distribution of probes along the genome, we chose to apply low stringent filters during the filtering process in order to keep the highest possible number of probes. Next, the selection step was done with the aim of selecting the best probe in all of the 150-base windows along *T. reesei* genome. This phase led to the selection of 222 690 oligonucleotides. Teolenn was able to design probes in 99.87% of all the possible 150-base windows along the *T. reesei* genome.

Comparison with other probe design software

We compared Teolenn results with some other tools available. To this end, the same *T. reesei* tiling design was done using ArrayDesign (7) and OligoTiler (4). Those solutions designed 144 201 (64,67%) and 222 778 (99,91%) probes, respectively. The first and the most striking result from this comparison is that, considering all designed probes with these three solutions, 4333 identical probes were found in two different designs and only 16 probes were common to the three designs.

We then compared the sensitivity of the probes obtained from each design. Figure 3 shows that Teolenn achieved a better combination of homogeneity and high median values regarding GC percent and T_m distribution. Indeed, Teolenn exhibited a median GC percent of 53.3 ± 0.07% instead of 53.3 ± 0.1% for OligoTiler and 46.6 ± 0.07% for ArrayDesign, and a median melting temperature of 74.3 ± 3.3°C instead of 74.3 ± 5.0°C for OligoTiler and 71.2 ± 3.3°C for ArrayDesign. It is noteworthy that no limitation was fixed during this calculation using Teolenn, contrary to ArrayDesign that exhibits no upper outlier (Figure 3A and B). In addition, we compared these results with the ones obtained using a variable probe length design. The T_m distribution of probes (Supplementary Figure S1) shows that we achieved only a slightly better T_m homogeneity median with variable probe length (74.35 ± 3.23°C) than with the fixed length design (74.29 ± 3.30°C).

The secondary structure minimum free energy tells us how much a probe is available for interaction. A low free energy creates strong secondary structures, which makes the probes unable to bind their target. As shown in Figure 3C, ArrayDesign obtained the narrowest and

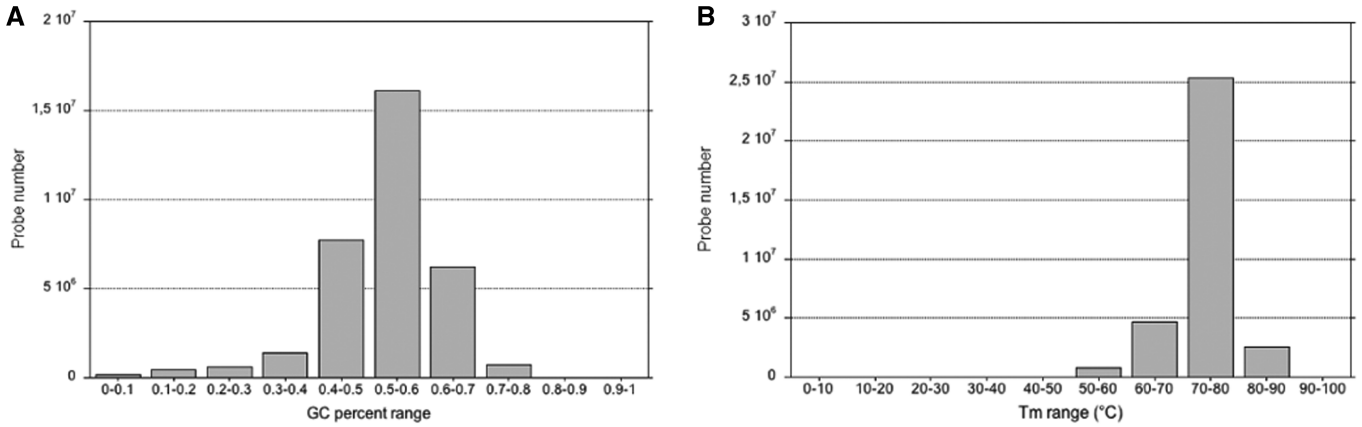

Figure 2. Library probes parameter distribution. Distribution of two parameters calculated for each probe of the oligonucleotide library. (**A**) Distribution of GC content for all possible probes from the library. The GC percent range is from 0 to 1 on the *x*-axis and the total number of probes in each category is shown on the *y*-axis. (**B**) Distribution of the melting temperature for all possible probes from the library. The T_m range in °C is displayed from 0 to 100 on the *x*-axis and the total number of probes in each category is shown on the *y*-axis.

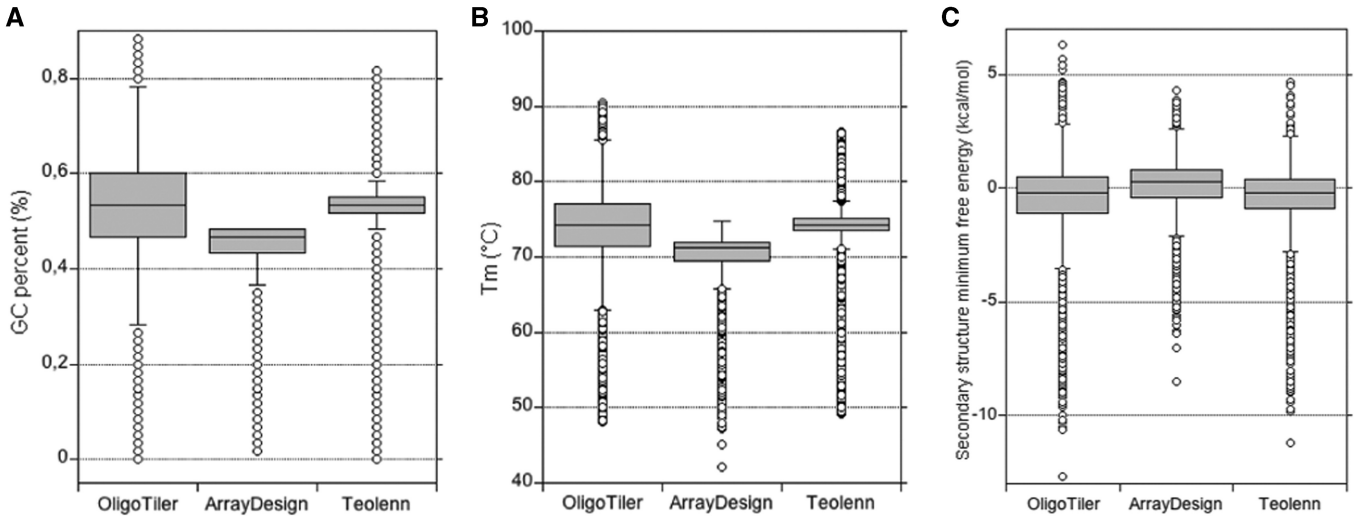

Figure 3. Comparison of the sensitivity of probe sets designed with OligoTiler, ArrayDesign and Teolenn software. For each probe set designed, boxplots show the distribution of (**A**) the GC percent, (**B**) the melting temperature (T_m) in °C and (**C**) the secondary structure free energy in kcal/mol.

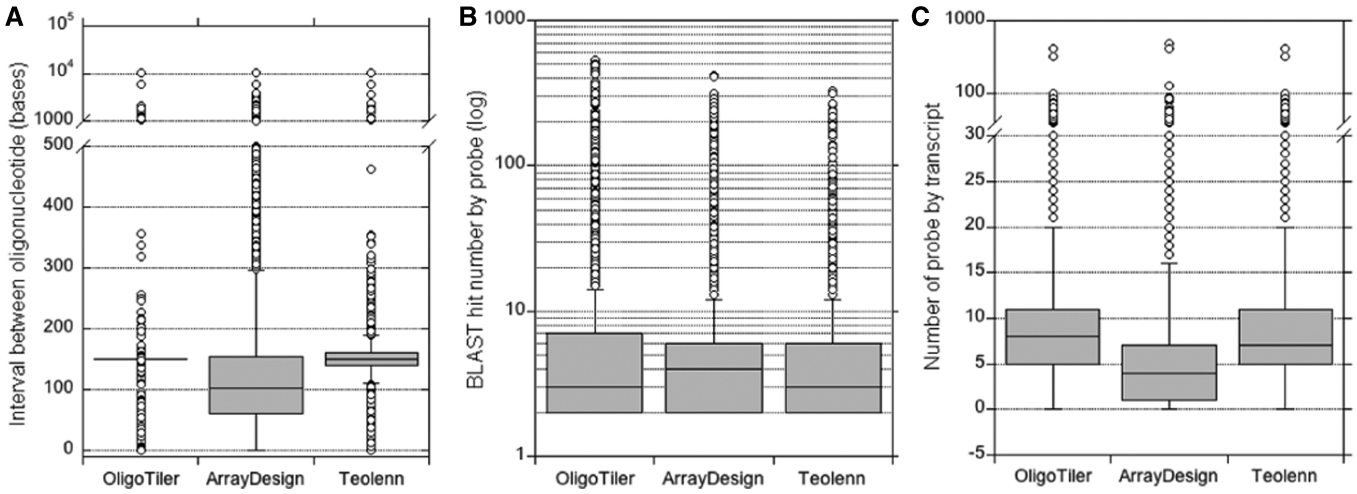

Figure 4. Comparison of the specificity of probe sets designed with OligoTiler, ArrayDesign and Teolenn software. (**A**) Distribution of the distance in base pairs between two consecutive oligonucleotides. (**B**) Distribution of the number of BLAST hits by oligonucleotide using the first Kane parameter (see 'Materials and Methods' section for details). (**C**) Distribution of the number of designed probes per annotated transcript in the reference genome.

highest free energy distribution of the three designed probe sets with a median of 0.3 kcal mol^{-1}. OligoTiler and Teolenn exhibited very similar distributions (median of –0.2 kcal mol^{-1}); Teolenn had a lower inter-quartile range (1.3 kcal mol^{-1}) than OligoTiler (1.6 kcal mol^{-1}).

We compared the respective distribution of the probes along the genome obtained with each design software (Figure 4A). Unsurprisingly, OligoTiler achieved the best even distribution of probes, since this program was created to optimize this parameter; the obtained median interval between two consecutive probes was 150 ± 27.8 bases. Teolenn almost reached the same value with a median interval of 150 ± 36.1 bases. This slight difference in standard deviation can be explained by the fact that Teolenn also optimizes probe specificity. Finally, ArrayDesign exhibited a 165 ± 221.3 bases median interval. This result is actually not surprising since ArrayDesign focuses on specificity optimization. This constraint leads to several windows with no probes selected, which artificially increases the interval between some probes.

The specificity of each probe set was investigated using the first Kane parameter (17). Using wuBLAST (12), we computed the number of hits for each probe that exhibited a full size alignment (60 bp) against the *T. reesei* genome with more than 75% of identity. The number of oligonucleotides with only one hit was slightly greater (97.1%) with Teolenn than with OligoTiler (96.8%) or ArrayDesign (95.8%). Considering only probes with more than one hit (Figure 4B), the median hit number by oligonucleotide was 4 with ArrayDesign and 3 with OligoTiler and Teolenn.

Finally, in order to estimate the coverage of potential biological entities, we measured the number of probes that cover transcript units for each design (Figure 4C). The annotation of *T. reesei* genome (16) lists a number of

9108 transcripts. We found that OligoTiler covered 98.9% of the annotated transcripts with 9.18 probes per transcript on average. Teolenn covered 98.6% of the annotated CDS with a mean of 9.10 probes per transcript, whereas ArrayDesign covered only 78.6% of the known transcripts (3.40 probes per transcript in average).

Analysis of a real hybridization

In order to assess whether our probe scoring strategy is able to reflect oligonucleotide quality, we wanted to evaluate the strength of the interaction between our designed probes and their targets within the context of a real hybridization. Indeed, some approximations (e.g. thermodynamic estimations) are made in order to calculate some of the scoring parameters. For example, the calculation of the free energy of the probe–target duplex refers to free molecules in solutions, whereas probes are linked to a slide surface with microarrays. To this end, we ordered the slide designed using Teolenn, and we performed a CGH using a self-hybridization of the *T. reesei* wild type reference strain (QM6a) against itself. In this context, no biological modifications are supposed to bias the observed results; this experiment is therefore very informative since the signal obtained for each probe is directly linked to the intensity of its interaction with its target. Image analysis and normalization were performed on a slide scan that exhibited no saturating spots, in order to get the highest number of exploitable measures.

We got a detectable signal (see 'Materials and Methods' section) for 97.4% of the spots on the array. We plotted the average intensity of each spot against its probe quality score. The linear correlation shown on Figure 5A clearly indicates a direct relationship between intensities and scores, even if the scatter of points is spread for high-scoring oligonucleotides ($R = 0.41$). This chart shows that, overall, our design strategy works properly, and that Teolenn is able to design high-quality probes. The intensity of scatter points found for oligonucleotide score over 0.6 (97.2% of detected probes) is representative of the classical bell shape distribution of signal intensities on microarrays; most intensities are found between 8 and 12. In addition, 95.8% of the detected features had good quality scores ($\geq$0.6) and intensities that were higher than background levels ($\geq$8). Oligonucleotides with low scores ($<$0.6, 2.8% of detected probes) mostly exhibited low intensities. Among them, only 901 had an intensity equal or greater than 10, which represents 13% of bad quality scores ($<$0.6) and 0.38% of all the detectable probes.

When looking at the distribution of the parameters that were calculated for these 'bad' probes, we found that all of them (i.e. T_m, GC percent, uniqueness and complexity) were lower than the ones associated with good oligonucleotides (Figure 5B). Finally, our score calculation was well adapted for 97.2% of the probes. Scoring calculation is also useful, because no linear correlation is detectable if the parameters are taken independently (Supplementary Figure S2).

DISCUSSION

Here we describe Teolenn, a new probe design workflow that was created to be flexible and customizable. To this end, we developed all the calculations, filtering and selection processes as independent modules. This organization is very useful for several reasons. First, the organization in modules allows for activating or disabling each function according to the available resources or to the needs of the user. For example, the complexity calculation can be deactivated if the masked reference genome of the organism is not available; in this case,

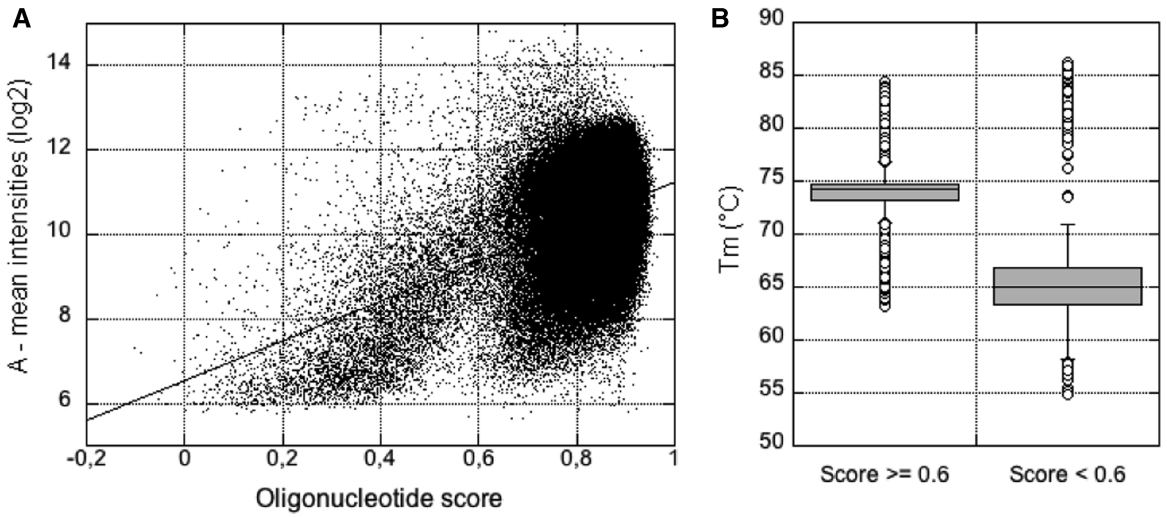

Figure 5. Correlation between oligonucleotide scores and spot intensities. (**A**) The graph displays for each probe the average of $\log_2$ intensities (A value) for the corresponding spot as a function of the oligonucleotide score calculated with Teolenn software. The straight line is the linear correlation between the two axes. (**B**) Distribution of T_m values for two sets of probes with intensities $>$10 (see A). Boxplots display the melting temperatures (T_m values) of high-quality scores ($\geq$0.6, left) and low quality ones ($<$0.6, right).

this parameter would not be taken into account for the calculation of the final probe quality score. Second, it is possible for developers to write new modules if needed. For example, the specificity calculation based on uniqueness can be replaced by another cross-hybridization calculation method like BLAST or suffix arrays (18). It is also possible to add modules for calculating the evaluation of secondary structures. Such calculation is time consuming; it is not implemented in the current version of Teolenn.

The flexible construction of Teolenn software fits every level of the probe design pipeline, particularly for the filtering and selection phases. The filtering process is done to influence the output of the probe design directly. Oligonucleotides can be filtered using probe quality. The probe library built at the beginning of the workflow contains all the possible non-redundant oligonucleotides. Without any filtering, a selection of the best probe is done in each window even in low-quality regions (e.g. repeats). This strategy is interesting for the users who want to get the most even distribution of probes along the genome. Conversely, it is possible to filter bad probes of the oligonucleotide library out. This ensures better results after hybridization but induces gaps in the tiling path, since no probes will be designed in AT rich or repeated regions, for example. Furthermore, the quality score and the calculated parameters are available for every probe designed by Teolenn. It is therefore possible to link bad detected signals with oligonucleotide scoring a posteriori, using the results obtained with control experiments (e.g. self-hybridizations). Here, we tested the influence of probe filtering on our *T. reesei* design. If we set up thresholds on T_m, GC percent and uniqueness in order to discard 15% of the worst probes, we were able to design probes in 98.45% of all possible windows compared with 99.87% without any filtering on probe quality. The probe-filtering step can also be used to filter probes according to genome annotations. Such filtering helps to create custom functional genomic microarrays in order to detect mRNA (transcriptome), miRNA, splicing events and so on. In these cases, the filtration step is done according to the coordinates of annotations along the genome. This filtration process can be combined, of course, with probe quality filtering and therefore allows for designing high-quality transcriptome arrays.

Customization can also be done with the probe selection process. Scores for each oligonucleotide are calculated during this last step, using the parameters obtained from Step 2 of the design workflow (Figure 1). The major advantage of customization at this stage is that users can set priorities among parameters. Indeed, a weight is assigned to each parameter in order to calculate the global oligonucleotide score. Changing these weights therefore modifies the way the best probe is selected in each window. For example, favouring the position parameter better than probe quality achieves a more even distribution of probes, but induces a wider distribution of sensitivity and specificity values. Setting the priority on melting temperature also makes it possible to obtain an almost isothermal distribution of the T_m in the probe design without modifying the length of each probe. In our example, introducing length flexibility in the design only led to marginal improvements, allowing an increase of the T_m median of 0.06°C and a decrease of 0.07 of the standard deviation.

Since it depends on the chosen window size, the position score is also calculated during the selection step. Here, the idea of 'window' can be interpreted in several ways. For a tiling microarray design, the window has to be a region with a fixed length all along the reference genome. In this case, the best position score is given to the most central oligonucleotide in each selected region. In contrast, for a custom array based on genome annotation (e.g. transcriptome arrays), the window size can be variable and dependant on coding sequences. The user may want to design a limited number of probes per transcript. In this case, the probe orientation is important, and selecting the most central one is not of interest. If only one probe is supposed to be designed, its position score is calculated in reference to either 5′- or 3′-end of the transcript, according to the reverse transcription method used (19). Teolenn is able to deal with all these cases for the design of probe sets whereas usual transcriptome probe design software only offer some of these choices (20–24).

All the parameters set by the user are gathered into one xml file, which allows convenient specification of the user properties during the process. However, developers have the possibility of directly modifying or adding new modules to the Teolenn workflow. Another worthwhile feature of Teolenn is that each step can be launched independently. Consequently, if users want to test the effects of different parameter weights on probe selection, Teolenn can launch only the selection step using results from the previous steps of the workflow, since they were already done. The calculation time for each test is therefore very short. It is noteworthy that the most time-consuming step, probe library construction, has to be done only once for each reference genome. Thus, Teolenn allows for testing parameters more easily than any other available probe design software. Furthermore, this way of working can also be useful for a large functional genomic project. With the same probe library, it is possible to first design a whole genome tiling array against a reference genome and then, when all transcript events have been detected without any a priori on gene annotation, to design a smaller and less expensive custom transcriptome array by filtering probes according to coding sequences. Thus, the flexibility of the Teolenn solution allows for quick and efficient construction of an infinite number of possible probe design solutions.

We designed probes against the genome of the fungus *T. reesei* with OligoTiler, ArrayDesign and Teolenn. Melting temperatures obtained with Teolenn were close to isothermal conditions with high median T_m, even when the oligonucleotide length was fixed. This was obtained without any cut-off on T_m values, contrary to ArrayDesign. Probes with a high T_m may cause saturating signals after scanning, whereas low T_m probes may lead to weak undetectable signals from the background. Since hybridization occurs at the same temperature for all probes during a microarray experiment, obtaining the narrowest T_m distribution helps avoiding low or high

signal detection. In addition, Teolenn is able to control simultaneously an even distribution of the tiling path. Since OligoTiler has been developed for tiling path optimization, Teolenn was not able to reach the same homogeneity level as this software, but it optimizes all sensitivity parameters at the same time, and without any loss of specificity. Indeed, Teolenn got the best result on specificity using the first Kane parameter; close to ArrayDesign, which was developed with the aim of getting the most specific probes along a tiling design.

With Teolenn, the user has permanent access to all the parameters and scores calculated during the probe design process. A posteriori correlations between probe quality score and signal strength on microarrays are therefore possible. Using a self-hybridization control array, we were able to detect a clear correlation between quality and signal intensities. This result validates our probe selection strategy. When investigating the properties of bad quality probes that exhibited a high signal level on our array, we were not able to understand this unexpected behaviour, since their parameters show low sensitivity and specificity for all parameters. This means that other properties that we were not able to evaluate may influence hybridization. However, these probes represented only 0.38% of all the detectable probes on the array.

Scores calculated by Teolenn may also be useful to validate some results after hybridization. Indeed, when working with transcriptome arrays, on which several probes are dedicated to the same transcript, discrepancies between probe results can sometimes be observed for the same transcript. This results could be explained either by a lack of information on transcript annotation (e.g. smaller RNA in the analysed condition, alternate splicing, incomplete reverse transcription resulting in 3′ truncated cDNAs), or by the fact that hybridization with low quality probes leads to bad signals. With Teolenn scoring values, it is now possible to differentiate between these two hypotheses.

In conclusion, we developed new flexible and customizable probe design software. Teolenn integrates a probe quality score calculation that is useful for correlations with signal strength after hybridization. The program is based on open modules in order to allow easy access to parameter setting and the development of new abilities. Teolenn probe design software is available at http://transcriptome.ens.fr/teolenn.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the ENS transcriptome platform staff Corinne Blugeon, Fanny Coulpier and Véronique Tanty for their support during hybridization experiments. They also thank Bernd Jagla (Pasteur Institute, Paris) for Teolenn testing on Mac OS.

FUNDING

Réseau National Génopôle (RNG); Infrastrutures en Biologie Santé et Agronomie (IBISA). Foundation Tuck 'Enerbio' fund doctorate fellowship to TP. Funding for open access charge: Work Institution.

Conflict of interest statement. None declared.

REFERENCES

1. Wold,B. and Myers,R.M. (2008) Sequence census methods for functional genomics. *Nat. Meth.*, **5**, 19–21.
2. Shendure,J. (2008) The beginning of the end for microarrays? *Nat. Meth.*, **5**, 585–587.
3. Lemoine,S., Combes,F. and Le Crom,S. (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Res.*, **37**, 1726–1739.
4. Bertone,P., Trifonov,V., Rozowsky,J.S., Schubert,F., Emanuelsson,O., Karro,J., Kao,M.Y., Snyder,M. and Gerstein,M. (2006) Design optimization methods for genomic DNA tiling arrays. *Genome Res.*, **16**, 271–281.
5. Lipson,D., Yakhini,Z. and Aumann,Y. (2007) Optimization of probe coverage for high-resolution oligonucleotide aCGH. *Bioinformatics*, **23**, e77–e83.
6. Schliep,A. and Krause,R. (2007) *Algorithms in Bioinformatics*, Vol. 4645. Springer Berlin, Heidelberg, Heidelberg, pp. 383–394.
7. Graf,S., Nielsen,F.G., Kurtz,S., Huynen,M.A., Birney,E., Stunnenberg,H. and Flicek,P. (2007) Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, **23**, i195–i204.
8. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
9. SantaLucia,J. Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
10. Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
11. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.
12. Chao,K.M., Pearson,W.R. and Miller,W. (1992) Aligning 2 sequences within a specified diagonal band. *Comp. Applications Biosci.*, **8**, 481–487.
13. Seidl,V., Gamauf,C., Druzhinina,I.S., Seiboth,B., Hartl,L. and Kubicek,C.P. (2008) The Hypocrea jecorina (Trichoderma reesei) hypercellulolytic mutant RUT C30 lacks a 85 kb (29 gene-encoding) region of the wild-type genome. *BMC Genomics*, **9**, 327.
14. Lemoine,S., Combes,F., Servant,N. and Le Crom,S. (2006) Goulphar: rapid access and expertise for standard two-color microarray normalization methods. *BMC Bioinformatics*, **7**, 467.
15. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
16. Martinez,D., Berka,R.M., Henrissat,B., Saloheimo,M., Arvas,M., Baker,S.E., Chapman,J., Chertkov,O., Coutinho,P.M., Cullen,D. *et al.* (2008) Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nat. Biotechnol.*, **26**, 553–560.
17. 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.
18. Manber,U. and Myers,G. (1993) Suffix arrays – a new method for online string searches. *SIAM J. Computing*, **22**, 935–948.
19. Tomiuk,S. and Hofmann,K. (2001) Microarray probe selection strategies. *Brief Bioinform.*, **2**, 329–340.
20. Bozdech,Z., Zhu,J., Joachimiak,M.P., Cohen,F.E., Pulliam,B. and DeRisi,J.L. (2003) Expression profiling of the schizont and

trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol.*, **4**, R9.
21. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
22. Nordberg,E.K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–1370.
23. 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.
24. Wernersson,R. and Nielsen,H.B. (2005) OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res.*, **33**, W611–W615.