Optimization of a Microarray Probe Design Focusing on the Minimization of Cross-hybridization

F. Horn[1], **H.-W. Nützmann**[2], **V. Schroeckh**[2], **R. Guthke**[1], **and C. Hummert**[1]
[1]Research Group Systems Biology / Bioinformatics
[2]Department of Molecular and Applied Microbiology
Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany

Abstract—*Microarrays are extensively used for high-throughput gene expression analyses in molecular biology. Microarray analysis is reliable if the probe binds specifically to the intended target transcript. Cross-hybridizations of microarray probes is one of the main systematic errors which is influenced by microarray probe design. Newly released genome annotations make it possible and necessary to improve given probe designs in order to reduce this source of error.*

We present a new method which evaluates and optimizes existing probe designs in a modular way. The workflow can include existing software and it can be adapted to additionally required probe design criteria. A microarray probe design optimization which focuses on the avoidance of cross-hybridization was exemplarily done for Aspergillus nidulans. We show the high impact of the underlying structural genome annotation on the probe design process. The new design was experimentally evaluated with the help of the mean variance of internal technical replicates.

Keywords: microarray, probe design, cross-hybridization, Aspergillus nidulans, optimization

1. Introduction

Microarray technique represents one of the most common methods to carry out genome-wide research based on sequenced genomes. A microarray experiment consists of many different steps which are all vulnerable to errors. Signal intensities strongly depend on the probe sequence, because different sequences generate varying physical properties, which are important for hybridization [1]. The properties of the probe sequences may be predicted and they are used for the microarray probe design [2].

The main objective of the design process is to increase the reliability of signal intensities by reducing systematic errors caused by the probe sequences. Among other criteria, the hybridization process itself is modeled with the help of criteria, like melting temperature uniformity, GC-content, prediction of secondary structures and Free Gibbs energy [3].

In order to guarantee a high discrimination between targets and non-targets, the probe design is checked for cross-hybridization. Cross-hybridization is a non-target binding between a probe and a transcript fragment which is not intended to match the probe. In fact, cross-hybridizations are one of the main sources of systematic error that affect tiling arrays [4] and even the well-established microarrays from Affymetrix [5]. Several studies have shown that nucleotide sequences are capable of hybridization, even when the complementary region between probe and transcript has only a 70% identity [1], [6]. Besides this identity threshold, non-specific bindings additionally need a longest continuous complementary substring of a certain minimum length [6], [3]. Signal intensities in the data may result from unspecific bindings and may lead to false-positively detected target genes.

There are approaches to cope with cross-hybridizations by creating new alternative Chip Definition Files (CDFs) of existing custom microarray probe designs [7], [8]. These methods correct and avoid the impact of cross-hybridizations by disregarding a certain fraction of the probes during data analysis. It is evident that the same level of information can be obtained with less probes spotted onto the microarray. The reannotation of oligonucleotide libraries is therefor the first step in order to obtain up-to-date microarray probe designs [9]. It is preferable to exclude existing cross-hybridizing oligonucleotides during the process of optimizing microarray probe designs [10]. This removal leads to a reduced production cost for each utilized data point. New alternative probes can be spotted onto the microarray which leads to a higher genome coverage rate or a higher number of replicates per gene.

Many different algorithms have been proposed for designing microarray probes [2]. Each algorithm has a different scope of application and consequently utilizes different probe design criteria and, as a consequence, perform differently. The different foci make it difficult to directly evaluate and compare the quality of the proposed algorithms with a theoretical optimization criterion. In fact, the limitations of the applied experimental protocol determine suitable probe design criteria and narrow down the set of available methods. It is favorable to use an extendable und adjustable general framework where different probe design criteria can be integrated [11], [12]. This allows to adjust for application-specific design criteria and enables the reuse of existing modular software.

In this work, we present a workflow which evaluates and optimizes an already given reference probe design concerning the avoidance of cross-hybridization. The optimization of the probe design is exemplarily done for a microarray for *Aspergillus nidulans* which is a model organism of filamentous fungi [13]. The obtained probe design minimizes unspecific bindings. We show that this design yields more reliable results. In addition to the avoidance of cross-hybridizations, it is possible to include different design criteria which are applied due to experimental constraints.

2. Results

2.1 Evaluation of reference probe design

The mapping of a given full-genome probe design for *Aspergillus nidulans* was examined by aligning the probe sequences against three structural genome annotations: two different versions available from the Broad institute and one version from the Central Aspergillus Data REpository (CADRE). (The annotations are referred to as BROAD (2008), BROAD (2010) and CADRE (2009), respectively.) For further information see methods and figure 1.

The given reference probe design contains 342 and 377 probes that cross-hybridize with BROAD (2008) and CADRE (2009) annotation, respectively (see table 1). Regarding the newer BROAD (2010) annotation, only 148 probes are considered as cross-hybridizing.

Using the BROAD (2008) annotation and the CADRE (2009) annotation respectively, 317 and 313 probes in the reference probe design do not match any transcript with a perfect sequence identity.

The reference probe design contains probes that do not match any transcript in the given annotation: 74 probes using BROAD (2008), 204 probes using CADRE (2009), and 993 probes using the newer BROAD (2010).

The reference probe design does not cover a number of predicted transcripts in each annotation: 442 transcripts in BROAD (2008), 478 transcripts in CADRE (2009), and as much as 968 transcripts in BROAD (2010).

The evaluation also calculated the thermodynamic properties of the probe sequences. The result reveals that the melting temperatures of the probes are in a narrow range between $80°C$ and $90°C$. This desirable property is achieved with the help of a uniform GC content of 48%.

In summary, the reference probe design is not optimized for any of the used annotations. Depending on the used annotation version, 7...11% of all probes do not match a transcript unambiguously. The current annotation causes a poorer performance which can be seen explicitly at the decreased number of perfect probes (see table 1).

2.2 Probe design optimization

A large fraction of the reference probe design is not optimized for any genome annotation and needs improvement. The objective of the optimization was to get 50 nucleotides long optimized oligonucleotides which use the BROAD (2008) annotation. The probes should be placed at the 5'-end because cDNA is used in the hybridization protocol.

The workflow of the proposed probe design method can be separated into three consecutive steps (see figure 2). In the first step new probe candidates are generated with the help of ArrayOligoSelector [14]. In the second step, probe candidates are evaluated with the help of evaluation tool to exclude cross-hybridizations (see above). The evaluation also calculates thermodynamic properties that are used in a following third step - a further selection. The selection step is necessary because only one probe sequence per gene is spotted.

The optimization showed that it was not possible to find a valid unique probe sequence for every transcript. In order to achieve a higher gene coverage, design criteria have to be mitigated. New probe candidates are iteratively generated from intervals of elongated transcript sequences. 1,303 probes were found in the smallest interval of 600 basepairs (see table 2). In the next two steps the interval is extended to 1,200 and 2,000 basepairs which only led to 30 and 24 additional probes, respectively. In a last step, probes that are capable of cross-hybridization are exceptionally allowed. The relaxation of this last criterion increased gene coverage with 53 additional probes. In total, the softening of the design criteria leads to 107 additionally covered genes in the presented study.

Finally, there are 188 genes without a valid probe sequence which leads to a transcript coverage rate of 98,2%.

The comparison of the resulting new probe design with the given reference probe design shows that the new probe design is optimized for the BROAD (2008) annotation (see table 1). The new design consists of 10,512 probes (99.5%) which match perfectly and do not show any cross-hybridization. Notably, the comparison with the reference probe design demonstrates that 254 genes are additionally covered in the optimized design while avoiding systematic errors.

Remarkably, there are also 214 extra covered genes if the CADRE (2009) annotation is used as basis. This result is achieved by a lower number of genes with systematic errors. The number of potentially cross-hybridizing probes is only 133 in comparison to 377 probes in the reference probe design. Only three specific probes match a transcript without a total sequence identity whereas this number is much higher in the reference probe design with 313 probes. Changes in the annotation lead to 143 probes that do not match any given transcript in contrast to 204 probes in the reference probe design. The number of uncovered genes is 264 which corresponds to a gene coverage rate of 97,5%.

For the current BROAD (2010) annotation the gene coverage of the probe design is reduced to 90,5% and the number of covered genes (9,561 vs. 9,592) is comparable between both versions of the probe design. Nevertheless, the new

Table 1: **Results of probe classification and gene coverage**

Annotation	BROAD (2008)		CADRE (2009)		BROAD (2010)	
Probe design	old	new	old	new	old	new
Number of probes	10,676	10,566	10,676	10,566	10,676	10,566
Perfect probes	9,943 (93.1%)	10,513 (99.5%)	9,782 (91.6%)	10,287 (97,4%)	9,535 (89.3%)	9,535 (90.2%)
Cross-hybridizing	342 (3.2%)	53 (0.5%)	377 (3.5%)	133 (1.3%)	148 (1.4%)	63 (0.6%)
Not identical match	317 (3.0%)	0 (0.0%)	313 (2.9%)	3 (0.0%)	0 (0.0%)	0 (0.0%)
Not matching	74 (0.7%)	0 (0.0%)	204 (1.9%)	143 (1.4%)	993 (9.3%)	968 (9.2%)
Total number of genes	10,701	10,701	10,546	10,546	10,560	10,560
Covered genes	10,259 (95,9%)	10,513 (98.2%)	10,068 (95.5%)	10,282 (97.5%)	9,592 (90.8%)	9,561 (90.5%)
Uncovered genes	442 (4.1%)	188 (1.8%)	478 (4.5%)	264 (2.5%)	968 (9.2%)	999 (9.5%)

Probes from the reference probe design (old) and the optimized (new) probe design have been mapped to different genome annotations. Probes either show no systematic error (perfect probes), hybridize with multiple genes (cross-hybridizing), match one gene without total sequence identity (not identical match), or do not match any transcript at all (not matching). The lower part of the table shows how many genes of the annotation are perfectly covered by the corresponding probe design.

Table 2: **Composition of the gene coverage**

	Number of genes
Reference probe design (validated probes)	9,103
Probe design optimization:	
Sequence range: 0…600 bp	1,303
Sequence range: 0…1,200 bp	30
Sequence range: 0…2,000 bp	24
Ignoring cross-hybridizations	53
Uncovered genes	188
Total	10,701

The gene coverage of the probe design results from different steps. A high number of genes are covered by validated probes from the reference probe design. The probe design optimization leads to an additional number of covered genes which are obtained by iteratively mitigating the probe design criteria. First, the transcript sequences are extended and at last the cross-hybridization criterion is relaxed. In the end, some genes remain that are not covered by any valid probe.

probe design still minimizes systematic errors. 63 probes are prone to cross-hybridizations in contrast to 148 probes in the reference probe design. A high number of 968 probes do not match any transcript at all which is again comparable to the performance of the reference probe design.

In summary, the new probe design reduces systematic errors regardless of the structural annotation used. Concerning the cross-hybridizations, the improvements become apparent. For BROAD (2008) and CADRE (2009) the gene coverage of the optimized probe design is higher as compared to the reference probe design.

2.3 Impact of genome annotation

The evaluation of different probe designs clearly highlights the big impact of the underlying structural genome annotation on the results (see table 1).

The new probe design was optimized for the BROAD (2008) annotation and the gene coverage could be increased to 98.2%. The optimization also takes effect for the CADRE (2009) annotation with a gene coverage rate of 97.5%. In comparison to the current BROAD (2010) annotation, the gene coverage rate is dramatically decreased to 90.5% which is comparable with the coverage rate of the reference probe design. The same trend for gene coverage can be seen for the reference probe design where the gene coverage rate also decreases to 90.8% if the BROAD (2010) annotation is used.

The differences in gene coverage result from probes which are vulnerable to systematic errors. The new probe design shows only a small fraction of probes that are prone to cross-hybridization in the BROAD (2008) annotation. This number doubles if the CADRE (2009) annotation is used. In the BROAD (2010) annotation only a few cross-hybridizing probes occur. This results from the increased number of error prone probes that do not match any transcript at all. The number of unmatched probes constitutes the largest error source which is affected by the change in genome annotation.

In the probe design optimized for BROAD (2008), the number of probes that are not classified as perfect increases from 54 (0.6%) over 279 (2,7%) to 1031 (9.8%) for the BROAD (2008), CADRE (2009), and BROAD (2010) annotation, respectively. The same trend holds for the non-perfect probes from the reference probe design which increases from 733 (6.9%) over 894 (8.4%) to 1141 (10.7%). It is noteworthy that a change in the annotation basis can cause almost 10% of all probes to be classified as invalid.

2.4 Experimental Validation

The new probe design is optimized for the minimization of systematic errors in respect to the BROAD (2008) annotation. Especially, the avoidance of cross-hybridization should significantly increase the reliability of experimental data. An indicator for improved reliability is a lower mean variance of internal technical replicates over each array. For this purpose, a highly reproducible experiment with the reference and the new probe design was performed (see methods). Microarray raw data was obtained from *Aspergillus nidulans* - *Streptomyces rapamycinicus* interaction experiments. The co-cultivation was performed because most of the secondary metabolite gene clusters are silent under laboratory conditions and the fungal-bacterial interaction leads to specific activations [15], [16]. (Microarray data is available at Gene Expression Omnibus - GSE25266.)

First, a microarray experiment using the reference probe design was performed. The following second experiment used the same experimental setup except that the new optimized probe design was used. It is not possible to compare the variance of probes for each single gene individually because an altered probe sequence has an essential impact on the signal intensities. Probes with the same nucleotide sequences have a high Pearson correlation coefficient of 0.928 whereas altered probe sequences result in a low correlation coefficient of 0.554.

Overall, the internal technical replicates should however show the desirable property of a lower mean variance over each array. The first experiment with the reference probe design used 4,148 internal technical replicates for 164 genes whereas the second experiment with the new probe design had 1,368 internal technical replicates for 157 genes. The mean variance of the internal technical replicates for the reference probe design range from 4.27…4.7 for the biological sample of the *A. nidulans*-*S. rapamycinicus* interaction and *A. nidulans* wildtype, respectively (see table 3). The new probe design shows a lower mean variance of internal replicates, namely 3.55 for the wildtype and 3.69 for the interaction sample. This change corresponds to an reduction of the mean variance with a ratio of 0.76…0.86. The application of a Shapiro-Wilk test indicated a normal distribution of signal intensities with a p-value < 0.05. An F-test with a subsequent Holm-correction confirmed the significance of the change in variance. All adjusted p-values are below 0.05. The lower mean variance over each array of the new probe design is significant. In summary, the statistical analysis of experimental results obtained from technical replicates supports the applied method and shows that the new probe design yields more reliable results.

Table 3: **Mean variance of technical replicates over each array**

Sample/Replicates	Old design	New design	ratio
A. nidulans rep1	4.79	3.73	0.78
A. nidulans rep2	4.69	3.85	0.82
A. nidulans mean	4.70	3.55	0.76
A. nidulans+*S. rapamycinicus* rep1	4.00	3.76	0.94
A. nidulans+*S. rapamycinicus* rep2	4.51	4.12	0.91
A. nidulans+*S. rapamycinicus* mean	4.27	3.69	0.86

Mean variance of internal technical replicates which were included in the first microarray experiment using the reference probe design and in the second experiment using the optimized probe design. Two technical replicates were used for each of the biological samples (*A. nidulans* and *A. nidulans* + *S. rapamycinicus*). Mean variances and the ratio between both experiments are given for each replicate and for the mean of each biological sample.

3. Discussion

3.1 Probe Design Optimization

The reliability of used probe designs need to be checked whenever new genome annotations are available [10], [9]. For *A. nidulans* the evaluation of the given reference probe design showed this necessity as it contains many systematic errors and the possibility to cover a higher number of transcripts is not fully exploited. The approach combines both steps - the evaluation of reference probe designs and the design of new probes. Frequently, a probe design already exists and probe sequences that satisfy the design criteria do not need to be recalculated.

It is challenging to find the right software which applies all probe design criteria described above. The usage of a modular workflow which allows for the flexible integration of different design criteria helps to adjust the oligonucleotide design to the specific experimental requirements. This approach allows the integration of own probe design criteria and existing software. A similar workflow with different steps has been proposed and implemented in the tool Teolenn [11]. This framework was not considered due to the missing integration of re-evaluation of existing probe designs.

For the generation of probe candidates many different software tools have been proposed. In the proposed workflow we decided to use ArrayOligoSelector [14] which applies a large fraction of required design criteria and was recommended in an evaluation of custom microarray applications [2]. The tool chosen is interchangable and should be orientated at the specific probe design requirements.

In this working example, hybridization are only considered if the alignment has a minimum sequence identity of 90% (see methods). This way, cross-hybridization can not be fully excluded because it was shown that it already occurs at a identity of 70% [6]. If the evaluation tool uses a more stringent cut-off, more probes are classified as invalid and more genes are not covered by any probe. The setting of this threshold is always a trade-off because the aim is to cover as many genes as possible while excluding cross-hybridizations. Hybridization with *S. rapamycinicus* transcripts was not checked because poly-dT-priming ensures that only eukaryotic RNA is amplified.

Due to the experimental objectives, the position of the probe and the GC content range were used as design criteria. The filtering for a narrow GC content range is a fast calculabe filter criterion and effectively obtains a close melting temperature uniformity. The computational costly application of the Nearest-Neighbor Model [17] gives a more precise estimation of the melting temperature. A direct application of this methods for probe design is limited because it assumes that both nucleotide strands interact freely in a solution which is not the case for microarrays.

Generally, if more probe design criteria are applied more probe candidates are excluded leading to a lower number of valid probe sequences. Overall, the used approach utilizes only a small set of all possible probe design criteria. Despite that, it was not possible to find a valid probe for 188 genes. Several factors contribute to this number of uncovered genes: If the gene annotation allows for transcripts which are shorter than the desired probe length or consist of highly repetitive sequence stretches, it is apparently not possible to find a valid probe sequence for them. In addition, a few transcripts share the same 3'-end, represent different splice variants, or are positioned within the same locus but on different strands. Finally, some sequences are at different loci, but have a high sequence similarity which may result from gene homology.

3.2 Impact of annotation databases

It is crucial to decide what structural genome annotation should be used as reference for the probe design. The reason are new genome assemblies and differences in the formal definition of the characteristics of a gene. Large fractions of the annotation of *Aspergillus nidulans* are done automatically with the help of bioinformatic tools. It is evident that with ongoing research the annotation of transcripts is subject to change. A large fraction of the oligonucleotide libraries can not be unambiguously matched to existing structural genome annotations [9]. The progress in laboratory research and, consequently, the related manual curation of genome annotations lead to more robust genome annotations.

3.3 Experimental Validation

The quality of the designed probes, and therefore the quality of the proposed approach, is eventually assessed by experimental validation. Probe sequences may be evaluated with spike-in experiments [18], self-hybridization experiments with the analysis of gene coverage [11], correlation of experimental data with probe design criteria [11], [12], experimental selection of probes [12], and the usage of internal technical replicates [19]. Without a transcriptome golden standard the impact of modifications can not be directly linked to the overall improvement of the array design. Spike-in experiments, Northern Blots, and qRT-PCR can only focus on a selection of chosen transcripts and are therefore not suited to assess a whole microarray probe design. Furthermore, it is not distinguishable which specific probe design criterion has an effect on the results because the criteria are mutually dependent. An altered probe sequence, for instance, does not only change the sequence similarity but also the physical properties of the probe and the hybridization. Nevertheless, it is necessary for an improvement of the design process.

In this study we used internal replicates to assess the quality of the new probes. Internal technical replicates allow to check for the performance of probes regardless of the experimental influences. A significant decrease of mean variances of internal replicates over each array was observed. This shows that the probes have a higher signal reproducibility. The optimized microarray probe design is more reliable as it has been shown with the help of statistically significant lower mean variance of the internal technical replicates.

4. Materials and Methods

4.1 Probe design evaluation

The probe design from febit biomed GmbH (Heidelberg, Germany) was as used as 'reference probe design' (see GSE25266 and [15]). It was analyzed regarding the structural genome annotations from BROAD institute [20] (two different versions downloaded October, 10th 2008 and February, 18th 2010) and from CADRE [21] (downloaded February, 16th 2009). The annotation versions are referred to as 'BROAD2008', 'BROAD2010', and 'CADRE2009', respectively.

Probe sequences were aligned locally to the known corresponding transcripts with the help of FASTA (Parameters: expectation value 1.0, alignment type 0) [22]. The thermodynamic properties of each probe and the hybridization were calculated with the nearest-neighbor model [17], which is implemented in the freely available software MELTING (Parameters: '-Hdnadna -N0.2 -P0.0001 -Ksan98a') [23]. A probe is considered to match a transcript if there is at least one 16 basepairs long common subsequence and if both sequences share a sequence identity not less than 90%. Although literature suggests that hybridization already occurs at 70% sequence identity [6], a less stringent cut-off was applied. A stricter constraint dramatically decreases the number of valid probe sequences and prevents a full-genome probe design. All probes are finally classified into four classes. Probes that i) match perfectly, ii) cross-hybridize, iii) do not match any transcript, and iv) hybridize, but are not fully identical with the target sequence.

4.2 Generation of new probe candidates

New probe candidates were generated for genes where no perfect matching probe is given in the reference probe design. Different available algorithms could be applied for this step. In this study, we integrated the public available tool ArrayOligoSelector (Parameters: target GC percentage 48.0, length of oligonucleotides 50), number of oligos per gene 5) [14], which utilizes sequence similarity, a given GC content range, tests for low-complexity regions, and recognition of self-complementary sequences. The transcript sequence was trimmed to the first 600 basepairs to reduce computational time and to meet the probe design objective of placing the probe near the 3'-end. The generated probe candidates were checked with the help of the evaluation tool described above. This guarantees that new probe candidates meet the given cross-hybridization criterion and that systematic errors are avoided.

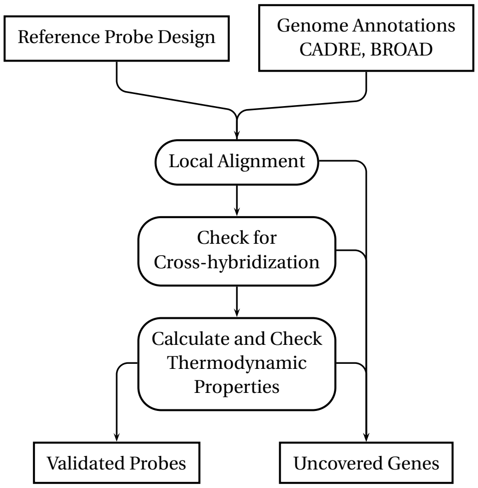
Reference Probe Design
Genome Annotations CADRE, BROAD
Local Alignment
Check for Cross-hybridization
Calculate and Check Thermodynamic Properties
Validated Probes
Uncovered Genes

Fig. 1: **Schematic overview of evaluation process.** A reference probe design is locally aligned to selected genome annotation databases. Probes that cross-hybridize are filtered and thermodynamic properties of the hybridization are calculated for further assessment.

4.3 Selection of validated probes

The aim of covering the full genome of *Aspergillus nidulans* allows only to spot one oligonucleotide for each gene considering the given spotting density constraint. Validated newly generated probe candidates are preferred if they are positioned at the 3'-end of the transcript. If several probes exist within an overlapping close interval of 50bp, the following second design criterion is applied: Probes with a GC content closest to the mean GC content of the reference probe design are chosen if the difference to the mean is below 8%. This ensures similar thermodynamic properties of all probes. After the application of these criteria, at most one single probe candidate per gene remains.

4.4 Iterative softening of design criteria

We start with a transcript sequence ranging from the 3'-end to 600 basepairs. In order to get a better gene coverage, the used transcript sequence range was iteratively extended to 1,200 and 2,000 basepairs for the remaining uncovered genes. Finally, the stringent cross-hybridization criterion was relaxed for the remaining uncovered genes. Hence, probe candidates are even considered if they are vulnerable to cross-hybridization. Probe sequences were chosen manually for genes of high biological interest and without a valid probe candidate. The manually chosen sequences minimize the number of cross-hybridizations and fall within the narrow range of the desired mean GC content (± 8%).

Merging the valid probes from the reference probe design with the selected new probe candidates resulted in the new and optimized probe design (see GSE25266 and figure 2).

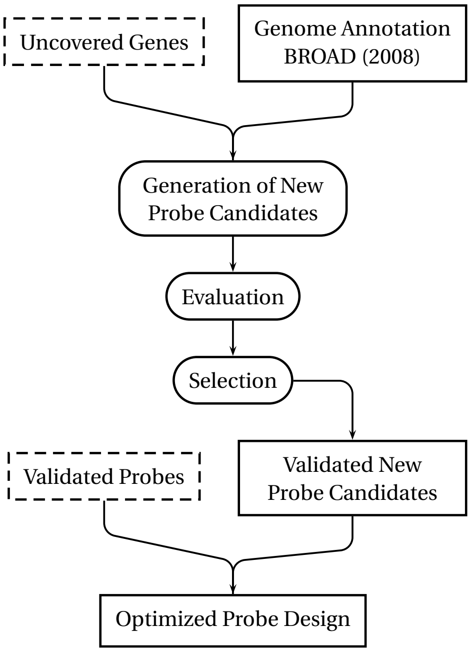
Uncovered Genes
Genome Annotation BROAD (2008)
Generation of New Probe Candidates
Evaluation
Selection
Validated Probes
Validated New Probe Candidates
Optimized Probe Design

Fig. 2: **Workflow of probe design optimization.** New probe candidates are generated for the genes where there are no current valid probe sequences. Probe candidates are evaluated with the evaluation tool. If more than one probe candidate is valid, different selection criteria are applied to select the best optimized probe. The final new optimized probe design is obtained by the combination of these probe candidates with the validated probes from the reference probe design. (Dashed lines represent results from the evaluation of the reference probe design.)

In summary, in this study the following probe design criteria have been applied: cross-hybridization, sequence complexity, lack of self-binding, GC content, and position on reverse strand.

4.5 Experimental validation

Microarray raw data was obtained from *Aspergillus nidulans - Streptomyces rapamycinicus* interaction experiments [15]. The fungus was incubated over night in liquid Aspergillus minimal medium (AMM) and shifted into fresh medium. *Actinomycetes* were cultivated in M79 medium and 5 ml of the culture was added to 100ml AMM and both organisms were further incubated at 37°C. The reference culture is incubated without bacteria. After 3 h, each sample was split into two identical technical replicates and total-RNA was isolated using RiboPure-Yeast Kit (Applied Biosystems) according to the manufacturers instructions. cDNA synthesis, labeling and microarray measurements were done by febit biomed GmbH. In the first experiment, the reference probe design was used. The same samples were used for the second experiment where the new probe

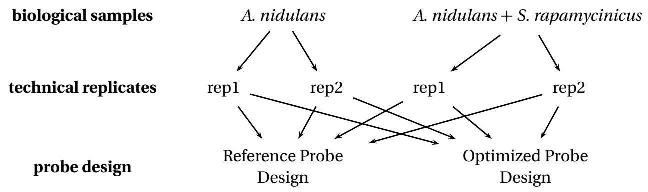

Fig. 3: **Schematic overview of Experimental Design.** In the first sample *A. nidulans* is cultivated without *S. rapamycinicus* and in the second sample it is co-cultivated with *S. rapamycinicus*. Each sample was split in two identical technical replicates. For each replicate a microarray experiment is performed with the reference and the new optimized probe design. The microarrays contain internal technical replicates that are used for the experimental validation.

design was utilized (see figure 3). All microarray data is compliant to the MIAME standard and can be accessed at GEO (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE25266.

Both microarrays contain several internal technical replicates which can be used to assess the quality of microarray design. The comparability of both experiments is shown with the help of Pearson correlation coefficients of the signal intensities. The mean variance of the internal technical replicates were calculated over each array. The application of a Shapiro-Wilk tests for a normal distribution of signal intensities. The significance of the change in variances are evaluated by an F-test and a subsequent Holm-correction.

5. Conclusion

We proposed a worflow for the evaluation and optimization of existing microarray probe designs. This workflow is capable of integrating existing software and adjusting the probe design according to the experimental requirements. Exemplarily, this approach has been applied for a full-genome microarray for *Aspergillus nidulans* with the focus on avoiding systematic errors, especially cross-hybridizations. The reduction of cross-hybridization improves the reliability of the probe design which can be seen in a reduced mean variance of internal technical replicates over each array. We showed the high influence of different structural genome annotations on the design process. It is recommended to check for cross-hybridizations based on a current version of genome annotation prior to microarray data analysis.

Acknowledgment

This work was supported by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS) and the Jena School for Microbial Communication (JSMC).

References

[1] T. R. Hughes *et al.*, “Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.” *Nat Biotechnol*, vol. 19, no. 4, pp. 342–347, Apr 2001.

[2] S. Lemoine *et al.*, “An evaluation of custom microarray applications: the oligonucleotide design challenge.” *Nucleic Acids Res*, vol. 37, no. 6, pp. 1726–1739, Apr 2009.

[3] Z. He *et al.*, “Empirical establishment of oligonucleotide probe design criteria.” *Appl Environ Microbiol*, vol. 71, no. 7, pp. 3753–3760, Jul 2005.

[4] S. Graf *et al.*, “Optimized design and assessment of whole genome tiling arrays.” *Bioinformatics*, vol. 23, no. 13, pp. i195–i204, Jul 2007.

[5] J. Mieczkowski *et al.*, “Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements.” *BMC Bioinformatics*, vol. 11, p. 104, 2010.

[6] M. Kane *et al.*, “Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays,” *Nucleic Acids Res.*, vol. 28, no. 22, pp. 4552–7, Nov 2000.

[7] F. Ferrari *et al.*, “Novel definition files for human GeneChips based on GeneAnnot.” *BMC Bioinformatics*, vol. 8, p. 446, 2007.

[8] M. Dai *et al.*, “Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.” *Nucleic Acids Res*, vol. 33, no. 20, p. e175, 2005.

[9] P. B. T. Neerincx *et al.*, “Oligorap - an oligo re-annotation pipeline to improve annotation and estimate target specificity.” *BMC Proc*, vol. 3 Suppl 4, p. S4, 2009.

[10] H.-H. Chou, “Shared probe design and existing microarray reanalysis using PICKY.” *BMC Bioinformatics*, vol. 11, p. 196, 2010.

[11] L. Jourdren *et al.*, “Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments.” *Nucleic Acids Res*, vol. 38, no. 10, p. e117, Jun 2010.

[12] F. Bidard *et al.*, “A general framework for optimization of probes for gene expression microarray and its application to the fungus Podospora anserina.” *BMC Res Notes*, vol. 3, p. 171, 2010.

[13] W. Vongsangnak and J. Nielsen, *Aspergillus: Molecular Biology and Genomics*. Caister Academic Press, Jan 2010, ch. Bioinformatics and Systems Biology of Aspergillus, pp. 61–84.

[14] Z. Bozdech *et al.*, “Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray.” *Genome Biol*, vol. 4, no. 2, p. R9, 2003.

[15] V. Schroeckh *et al.*, “Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in Aspergillus nidulans.” *Proc Natl Acad Sci U S A*, vol. 106, no. 34, pp. 14 558–14 563, Aug 2009.

[16] A. A. Brakhage and V. Schroeckh, “Fungal secondary metabolites - strategies to activate silent gene clusters.” *Fungal Genet Biol*, Apr 2010.

[17] J. SantaLucia and D. Hicks, “The thermodynamics of DNA structural motifs.” *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 415–440, 2004.

[18] I. V. Yang, “Use of external controls in microarray experiments.” *Methods Enzymol*, vol. 411, pp. 50–63, 2006.

[19] D. L. Leiske *et al.*, “A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray.” *BMC Genomics*, vol. 7, p. 72, 2006.

[20] J. E. Galagan *et al.*, “Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae.” *Nature*, vol. 438, no. 7071, pp. 1105–1115, Dec 2005.

[21] J. E. Mabey *et al.*, “Cadre: the Central Aspergillus Data REpository.” *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D401–D405, Jan 2004.

[22] W. R. Pearson, “Flexible sequence similarity searching with the FASTA3 program package.” *Methods Mol Biol*, vol. 132, pp. 185–219, 2000.

[23] N. L. Novère, “MELTING, computing the melting temperature of nucleic acid duplex.” *Bioinformatics*, vol. 17, no. 12, pp. 1226–1227, Dec 2001.