Bioinformatics, 32, 2016, i552–i558
doi: 10.1093/bioinformatics/btw451
ECCB 2016

OXFORD

Optimization of signal-to-noise ratio for efficient microarray probe design

Olga V. Matveeva[1,2,*], Yury D. Nechipurenko[2], Evgeniy Riabenko[3], Chikako Ragan[4], Nafisa N. Nazipova[5], Aleksey Y. Ogurtsov[6] and Svetlana A. Shabalina[6,*]

[1]Biopolymer Design LLC, Acton, MA 01721, USA, [2]Engelhardt Institute of Molecular Biology, Moscow 119991, Russia, [3]Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, 141701, Russia, [4]Queensland Brain Institute, University of Queensland, Brisbane, QLD 4072 Australia, [5]Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, 142290, Russia and [6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Target-specific hybridization depends on oligo-probe characteristics that improve hybridization specificity and minimize genome-wide cross-hybridization. Interplay between specific hybridization and genome-wide cross-hybridization has been insufficiently studied, despite its crucial role in efficient probe design and in data analysis.

Results: In this study, we defined hybridization specificity as a ratio between oligo target-specific hybridization and oligo genome-wide cross-hybridization. A microarray database, derived from the Genomic Comparison Hybridization (GCH) experiment and performed using the Affymetrix platform, contains two different types of probes. The first type of oligo-probes does not have a specific target on the genome and their hybridization signals are derived from genome-wide cross-hybridization alone. The second type includes oligonucleotides that have a specific target on the genomic DNA and their signals are derived from specific and cross-hybridization components combined together in a total signal. A comparative analysis of hybridization specificity of oligo-probes, as well as their nucleotide sequences and thermodynamic features was performed on the database. The comparison has revealed that hybridization specificity was negatively affected by low stability of the fully-paired oligo-target duplex, stable probe self-folding, G-rich content, including GGG motifs, low sequence complexity and nucleotide composition symmetry.

Conclusion: Filtering out the probes with defined 'negative' characteristics significantly increases specific hybridization and dramatically decreasing genome-wide cross-hybridization. Selected oligo-probes have two times higher hybridization specificity on average, compared to the probes that were filtered from the analysis by applying suggested cutoff thresholds to the described parameters. A new approach for efficient oligo-probe design is described in our study.

Contact: shabalin@ncbi.nlm.nih.gov or olga.matveeva@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Oligo-probes are involved in hybridization with their specific DNA or RNA targets in many biotechnology applications; one such application is microarray technology. Microarrays are reliable and currently more cost effective than RNA-Seq for gene expression profiling in model organisms. Moreover, RNA-Seq is not widely used for Comparative Genomic Hybridization for gene copy number evaluation due to prohibitive costs. RNA-Seq will eventually surpass microarray for routine use, but currently the techniques can be complementary to each other.

The focus of this study is in improvement of oligo-probe design for microarray technology. The main problem of any microarray experiment is that the specific and efficient oligo-target duplex formations are usually combined with non-specific parallel reactions. In addition to specific oligo-target duplex formation, which represents interaction between an oligo-probe and a fully complemented target,

an oligo-probe could interact with many partially complemented DNA or RNA sequences. These interactions lead to the formation of many non-perfect duplexes, which are responsible for cross-hybridization signals. Other types of non-specific reactions include oligo-probe self-folding, target intra- and inter-molecular interactions. Microarray hybridization data are a great resource for characterizing these parallel reactions. A single microarray experiment, especially with comparative genomic hybridization, allows visualization of thousands or even millions of hybridization reactions. Such experiments represent 'magnifying glasses' for observing enormous diversity in hybridization behavior of oligo-probes. No other technology provides such a high volume of useful information for analysis of specific oligo-target hybridization in a complex mixture of different non-specific reactions (Chudin *et al.*, 2002; McCall and Irizarry, 2008; Zhang *et al.*, 2003). Evaluation of a probe's ability for cross-hybridization is very important for optimization of oligo-probe design. The existence of non-Watson-Crick base pairings between a probe and genomic DNA, specifically, for G-rich probes and particularly for those probes with certain G motifs, complicates hybridization data analysis and efficient oligo-probes selection procedures for array design (Binder *et al.*, 2009; Fasold *et al.*, 2010; Langdon *et al.*, 2008; Memon *et al.*, 2010; Wu *et al.*, 2007; Upton *et al.*, 2008; Matveeva *et al.*, 2003). Local nucleotide profiling has revealed that probes with G-rich sequences at their 5′ ends are more vulnerable to a hybridization 'G negative' effect than probes with G-rich sequences at their 3′ ends (Binder *et al.*, 2005).

Human Affymetrix microarray datasets that included 'empty' probes, where complementary targets for these probes are absent from the human genome, are very useful for cross-hybridization studies. Consequently, the signals from 'empty' probes are caused by cross-hybridization only. The analysis of these types of databases contributed to a creation of an advanced model for cross-hybridization prediction (Furusawa *et al.*, 2009).

Complex connections between hybridization specificity and genome-wide cross-hybridization have not been sufficiently studied for optimal oligo-probe design. It is widely accepted, but not always true, that probes with the lowest cross-hybridization signal (low noise) are most specific (have highest hybridization specificity). Low cross-hybridization signal is not necessarily indicative of the hybridization high specificity (high signal). Low cross-hybridization signal could be a result of the poor ability of the probes to interact in general; in such cases, both signal and noise are low. Conversely, probes that generate high specific hybridization can also generate high cross-hybridization.

With some exceptions, it is still largely unknown how probe characteristics such as oligo-target duplex stability, probe's self-folding or nucleotide content can influence signal-to-noise ratios in array experiments. More specific probes could be designed if these influences are identified and characterized.

The calculation of the Gibbs free energy change (ΔG), which accompanies the hybridization reaction, can be performed in a number of different ways. It has been shown that the ΔG of the binding reaction, calculated as the sum of derived nearest neighbor parameters, obtained from solution studies, to a certain extent correlates with array signal intensity (Weckx *et al.*, 2007; Wei *et al.*, 2008; Xia *et al.*, 2010; Zhang *et al.*, 2007). However, ΔG of the direct reaction calculated as the sum of position-dependent weighted nearest neighbor parameters correlates with array signal intensity better (Zhang *et al.*, 2003). Correlation was also improved when the parameters were calculated directly from array experiments (Hooyberghs *et al.*, 2009; Zhang *et al.*, 2003). Despite the parameter investigations referenced in the above studies, a universally acceptable set of nearest neighbor parameters for microarray hybridization has not yet been established.

In this study, we investigated how the specificity of hybridization, which is defined as the ratio between specific and cross-hybridization, is affected by the probe characteristics mentioned above. For this purpose, 'empty' and 'full' probes, with targets located on the human X chromosome, were included in a dataset, derived from an Affymetrix tiling microarray Comparative Genomic Hybridization (GCH). Total hybridization consists of two components: specific and cross-hybridization. It is not possible to split the signal of each individual probe into these components for the hybridization data. Nevertheless, the approach we call 'binning and averaging' allowed analysis of probes' features and comparison of the probe signals from each bin. This approach promoted the investigation of relationships between probes' theoretical characteristics and experimental hybridization specificity values. Pre-filtering and removing of oligonucleotides that are capable of forming canonical or non-canonical secondary structures or probes with low sequence complexity and asymmetry altogether improves this categorization.

A genomic cross-hybridization signal is caused by reactions between an oligo-probe and multiple partially complementary sequences in a genome. We investigated whether computational software, which evaluates duplex stability between an oligo-probe and partially complemented targets, can be useful for modeling and predicting the hybridization specificity of probes. The software packages that we tested are 'FASTH' (Ragan *et al.*, 2009), 'OligoArrayAux', which is a subset of the UNAFold package (Rouillard *et al.*, 2003), 'NCBI-Hybrid' (Matveeva and Shabalina, 1993; Shabalina 2002; Shabalina *et al.*, 2006; Matveeva *et al.*, 2007) and 'Osaka University Software' (Furusawa *et al.*, 2009). All these programs are created for different research purposes: 'FASTH' and 'NCBI-Hybrid' are created for transcriptome-wide search of target candidates for microRNAs, 'OligoArrayAux' — for genome-scale oligonucleotide microarray design, and 'Osaka University Software' — for microarray oligo-cross-hybridization evaluations.

Despite the differences, all programs are able to output values that can be mathematically transformed into hybridization specificity evaluations using basic equations of equilibrium thermodynamics. In summary, our approach shows that even though duplex stabilities between oligo-probes and partially paired sequences can be used for modeling and predicting probes' hybridization specificity, these calculations are comparable with the simplified approach based on the duplex stability between fully complemented probes. The oligo-target duplex stability calculation procedure is more transparent and less time consuming. Duplex stabilities between an oligo-probe and fully complemented target could be significantly discriminative between highly specific and non-specific hybridization probes. In addition, we showed that stable probe self-folding, high G-rich content, presence of GGG motifs, low sequence complexity and symmetry are all parameters that diminish hybridization specificity.

2 Materials and methods

2.1 Hybridization database of 25-mers obtained after the GCH experiment

In normal human somatic chromosomes, each gene is represented by two copies (Supplementary Fig. S1A). In the male X chromosome, a majority of genes is represented by one copy (Supplementary Fig. S1B). In male patients affected by Duchenne muscular dystrophy syndrome (DMD), the region of the *DMD* gene is deleted and

consequently represented by zero copies (Supplementary Fig. S1). We analyzed the hybridization data obtained from an experiment performed with DNA from a DMD syndrome male patient, where a large part of the *DMD* gene in the X chromosome was deleted in the patient's DNA (Supplementary Fig. S1C). Consequently, the oligo-probes targeting the deletion region of the *DMD* gene are 'empty'; they correspond to the background of genomic cross-hybridization signals. While oligo-probes that target the non-deleted region in the X chromosome are 'full'; they correspond to the sum of specific and cross-hybridization signals. This sum is referred to as the 'total hybridization' in the study. Both sets of 'empty' and 'full' probes include 10 000 data points from the same hybridization experiment, performed on the same chip. The standard Affymetrix protocol was used for genomic DNA amplification and hybridization at 45°C. Hybridization was performed using a tiling array GeneChip Human Mapping 100K Set.

2.2 Oligo-probe sequence characteristics

The list of sequence characteristics studied in this work includes probe nucleotide content, sequence complexity, ability to form stable secondary structures or self-folding and ability to form stable oligo-target duplexes. In addition, we also considered how the position of some mono-, di- and tri-nucleotides in a probe sequence affects probe hybridization efficiency using in-house scripts (Kondrashov and Shabalina, 2002; Ogurtsov *et al.*, 2008; Webb *et al.*, 2002). The ability of probes to be self-folded and to form stable oligo-target duplexes was evaluated by calculating the ΔG of relevant reactions. ΔG value of oligo-probe secondary structure was calculated by our Afold software (Ogurtsov *et al.*, 2006) and by the DINAMelt program (unafold.rna.albany.edu/?q = dinamelt). ΔG values of oligo-target duplexes were calculated by a in-house script. The ΔG values were evaluated using previously published nearest neighbor parameters (SantaLucia *et al.*, 1996).

The sequence asymmetry and simplicity (SAS) score is a measurement of asymmetry between A-T and G–C frequencies of the given word (string), calculated as a sum of differences between A and T nucleotides, and between G and C nucleotides—$(\%A–\%T)^2 + (\%G–\%C)^2$. In order to find a value that is proportional to sequence complexity, we introduced the 'Symmetric Complexity' or SC score. Its value was calculated as $SC = 1 - SAS/SAS_{max}$, where SAS_{max} is the maximum possible value of sequence simplicity, which is characterized by an oligo-probe consisting of a single nucleotide and equal to 1(see details in Supplementary Materials).

2.3 Relationships between hybridization intensities and probe sequence characteristics

The hybridization data were separated into bins according to their physical or sequence characteristics, such as perfect oligo-target duplex stability or particular nucleotide count, and the average hybridization signals were computed in each bin. Binning and averaging involves separating the probes into bins according to their sequence characteristics and computing the average signal in each bin. The difference between the average signal from a bin with 'empty' probe (cross-hybridization) and a 'full' probe bin (total hybridization) can represent the average signal of specific hybridization for all probes characterized by similar characteristics. The calculation of this difference can be used for evaluation of the signal-to-noise ratio between specific- and cross-hybridization, i.e. the average specificity of hybridization for each probe's bin. Modeling theoretical cross-hybridization and calculation hybridization specificity was based on the estimation of occupancy distribution of molecules on target DNA (Landau and Lifshitz, 1980; Nechipurenko, 2015; Segal and Widom, 2009; see Supplementary Materials). Hybridization specificity was calculated as the ratio between specific- and cross-hybridization. To evaluate the distribution of hybridization specificity inside each bin, bootstrap analysis was performed (see Supplementary Materials for details).

3 Results

3.1 Filtering probes with negative characteristics

We categorized the probes in both datasets according to free energy of fully paired duplex, their secondary structure stabilities, nucleotide content, presence of some sequence motives, predicted hybridization affinity, hybridization specificity and some additional factors. We found that some of these factors negatively affect probes' hybridization specificity. These factors were called 'negative probe characteristics'. We found that such probe characteristics include (i) probes' ability to form canonical secondary structure, (ii) high G content, (iii) presence of at least one GGG motive and (iv) lack of sequence asymmetry (specifically, G versus C) and complexity. We developed an approach for the calculation of probes' sequence asymmetry and complexity, referred to as the SC score (see Materials and Methods). Each step of filtering was characterized by at least one type of negative characteristic, which improved hybridization specificity (Fig. 1). All filtration steps together accomplished one after another have a synergetic cumulative effect.

3.1.1 Secondary structure of probes diminishes hybridization specificity

We found that high secondary structure stability affects probes' hybridization specificity. Higher stability corresponds to lower specificity (Fig. 1A). We suggest avoiding oligos with self-folding potential for optimal probe design.

3.1.2 G-rich probes have low hybridization specificity

We found that G-richness affects probes' hybridization specificity in a negative way (Fig. 1B). Because of this finding, we suggest avoiding oligos with G content equal or greater than 8 nucleotides. We found that the presence of at least one GGG block affects probe's hybridization specificity in a negative way (Fig. 1C). Taking into account the results of searching for G-blocks, we suggest that oligos with at least one GGG block should be excluded from analysis. Both G-richness and the presence of GGG motifs in the probes synergistically decrease the hybridization specificity.

3.1.3 Symmetric sequence complexity is associated with high hybridization specificity

We have found that hybridization specificity and the SC score correlate positively, such that hybridization specificity increases along with an increase of SC (Fig. 1D). In spite of the exclusion of G-rich probes from the analysis, a significant increase of hybridization specificity was demonstrated by the categorization procedure according to SC score.

3.1.4 Probes that form unstable duplexes with their targets have low hybridization specificity

We discovered that probes that form the least stable duplexes are least specific, despite the fact that a cross hybridization component of the signal for these probes is also small (Fig. 2, left panels). Additionally, the probes that are forming most stable duplexes are

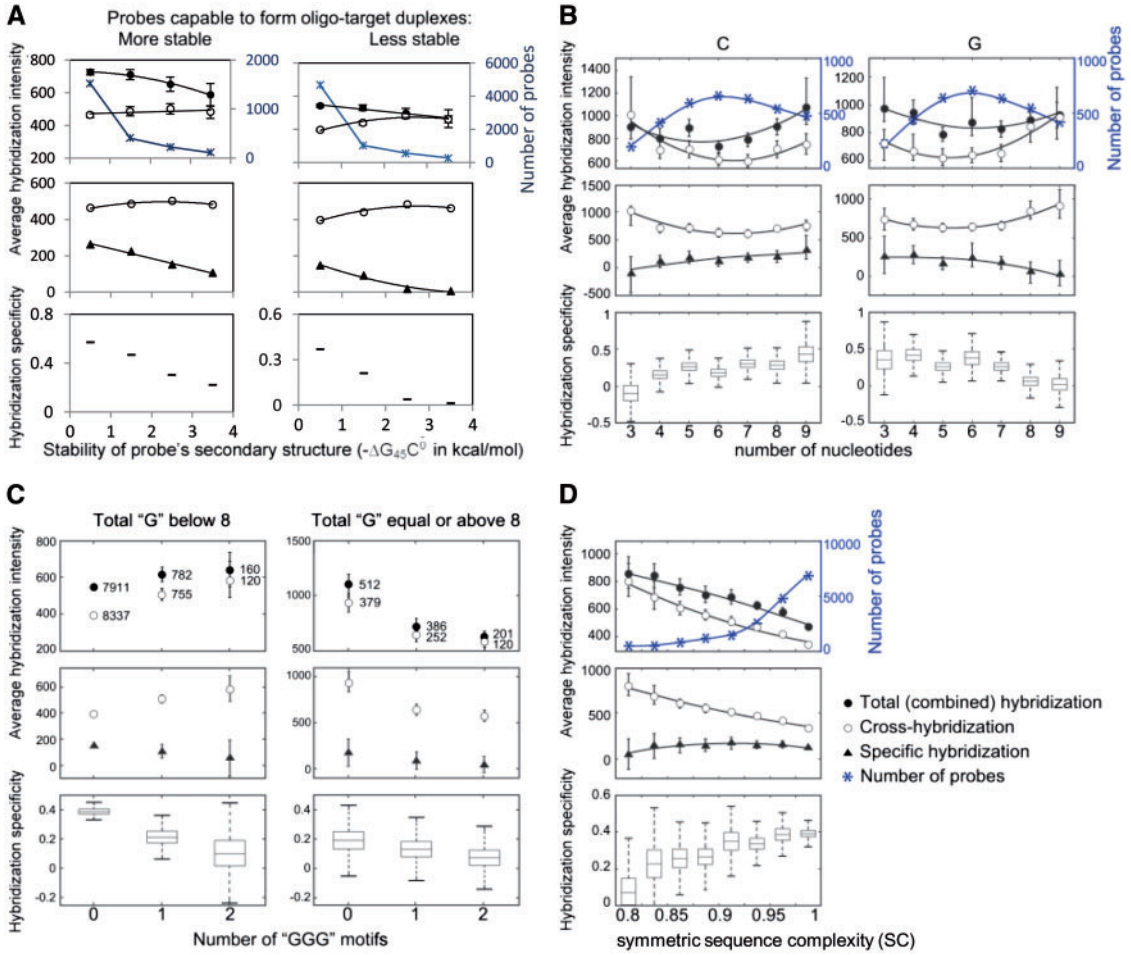

Fig. 1. Relationship between hybridization and oligo-probes thermodynamic and sequence characteristics. The averaged signals from each bin correspond to total (combined) hybridization if these values were calculated using the 'full' probes. The difference between total and cross-hybridization corresponds to specific hybridization. The numbers of probes in each bin are indicated numerically or shown as connected blue stars using the secondary axis. (**A**) The probes in each category (panel) were separated into bins according to the ability of the probe to form stable secondary structures. The left panel shows the plots of probes, which are able to form more stable oligo-target duplexes (ΔG below or equal to −26 kcal/mol), and the right panel shows plots of probes, which form less stable duplexes (ΔG above—26 kcal/mol). (**B**) The probes in each category (panel) were separated into bins according to number of nucleotides in each probe and an average hybridization signal was calculated for each bin. The left panel shows plots created with C, while the right one shows plots created with G nucleotide. Only scatter plots for probes which form stable duplexes with their targets (ΔG below—26 kcal/mol) are shown. (**C**) The probes in each category (panel) were further separated into bins according to the number of present GGG motifs and average hybridization intensities were calculated for each bin. The left panel shows plots created for probes containing 7 or less G. The right panel shows plots created for probes containing 8 or more G. (**D**) The probes were separated into bins according to SC score and average hybridization intensities were calculated for each bin

the most specific ones, despite having comparatively high cross-hybridization values.

3.2 Theoretical hybridization specificity

Stabilities of partially complemented duplexes (ΔG_{cross}) and accumulated cross-hybridization signals were estimated by 'NCBI-Hybrid', 'FASTH' and 'OligoArrayAux' using complete nucleotide sequences of human chromosomes as described in the Supplementary Materials. We calculated theoretical hybridization specificity of each probe in the database by processing different software outputs with predicted partially complemented oligo-target duplexes (see Supplementary Materials). Using a binning and averaging approach, the distribution of hybridization specificity predicted by the 'NCBI-Hybrid' program is shown on Figure 2 (right panels). The analysis revealed that experimental hybridization specificity is growing along with the growth of theoretical hybridization specificity. For comparison, a binning and averaging approach was applied to categorization of probes by oligo-target duplex stability (Fig. 2, left panels).

The discovered relationships help in categorization of probes into N subsets that are different in hybridization specificity. The probes predicted to be least specific have specificity values ranging from 0 to 0.3. The probes predicted to be more specific have specificity values ranging from 0.4 to 0.6. We found that categorization according to theoretical specificity values allows the detection of most specific oligo-probes bins with specificity value ranging from 0.4 to 0.6, regardless of the software that we used for prediction of stabilities between oligo-probes and their partially paired sequences (Fig. 2 and Fig. 3).

Approximately 10–20% of the probes from the testing database have hybridization specificity above 0.45. The majority of probes in the database have specificity values of at least two times less, with a median value of ∼0.23. Thus, probe selection procedures described in this study can be helpful for detection of the most specific hybridization probes with a specificity of at least two times higher than that of the majority of remaining probes. Comparison of the categorization results for 'NCBI-Hybrid' output and categorization

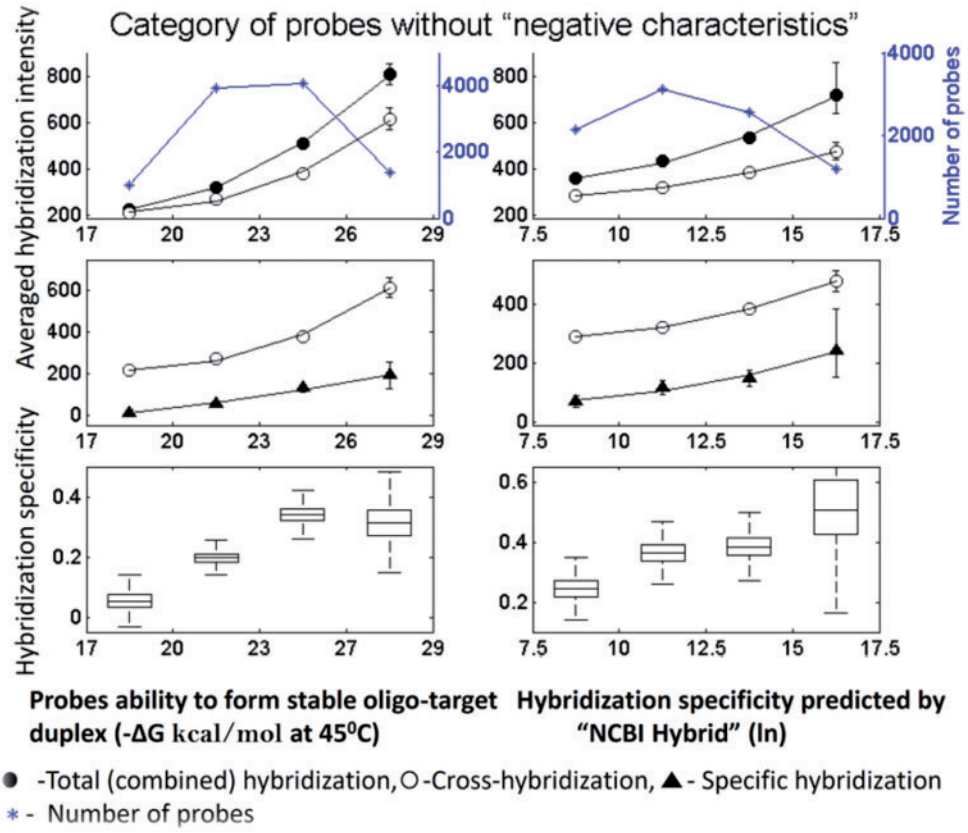

Fig. 2. Relationship between hybridization and oligo-target duplex stability. The data were separated into bins according to probes' ability to form a stable duplex with its target (left panel) or according to 'NCBI-Hybrid' predicted hybridization specificity (right panel). Hybridization specificity was calculated as the ratio between specific and cross-hybridization using bootstrap with 10 000 resamplings. The estimates of hybridization specificity in each bin are presented as box plots denoting a median, 25th and 75th percentiles, and whiskers to the most extreme data points not considered outliers. Only a category of the probes without negative characteristics is shown. These negative characteristics include low sequence complexity (SC score below 0.95), high amount of G (above 7), presence of GGG motif and stable secondary structures ($\Delta G \leq -2$ kcal/mol)

according to the oligo-target duplex stability (Fig. 2) showed that the results of these two approaches are comparable. Thus, both of these approaches are efficient for oligo-probe design; however, the oligo-target duplex stability calculation procedure is more transparent and less time consuming.

3.3 Cross-hybridization modeling and calculation of hybridization specificity by different methods

An investigation of the relationship between experimental and predicted hybridization specificity, estimated by three different programs, was performed among all probes and among four probes' subsets after filtrations (Fig. 3). The probes were filtered step by step removing oligos possessing one, two or more 'negative characteristics' mentioned in the previous sections. For performing analysis, data were separated into bins according to predicted hybridization specificity. After this, the experimental hybridization specificity was evaluated in each bin (Fig. 3). We discovered a relationship between theoretical and experimental hybridization specificity, which is stronger among probes without negative characteristics. An example of such a relationship for all these approaches is shown in Figure 3.

Figure 2 demonstrates that both specific and cross-hybridization increase along with the growth of predicted hybridization specificity (top and middle plot). Experimental hybridization specificity also increased along with the growth of theoretical hybridization specificity (right panel, bottom plot). The category of most specific oligos selected from the probes without 'negative characteristics' has the experimental hybridization specificity value of ~0.5 for 'NCBI-Hybrid'. A similar specificity value or even higher was reached by all other programs that were used in this study (Fig. 3). Thus, filtering out the probes with 'negative characteristics' improves the relationship between experimental and theoretical hybridization specificity values in all predictions.

Taking into account that the majority of probes in the database have a specificity value close to 0.23, the application of probe selection procedures described in this study could double hybridization specificity of micro-arrayed oligo-probes.

Filtering out the probes with different combinations of 'negative characteristics' has a cumulative effect on the increase of hybridization specificity among probes with high predicted hybridization specificity (Fig. 3, the bins marked with the percentage of included probes). The effect is visible for probes separated into bins according to probes' ability to form a stable duplex with its target or according to software predicted hybridization specificity. Regardless of software used for hybridization specificity prediction, each filtration step increases the average hybridization specificity of the probes.

The binning and averaging approach does not allow creation of a continuous function to characterize a relationship between experimental and theoretical values. Since the approach does not allow estimation of a standard deviation error corridor in this continuous function, we employed bootstrap analysis to overcome these limitations (Supplementary Materials; Fig. 3). This analysis has shown that in spite of the higher level of variability in the last bin,

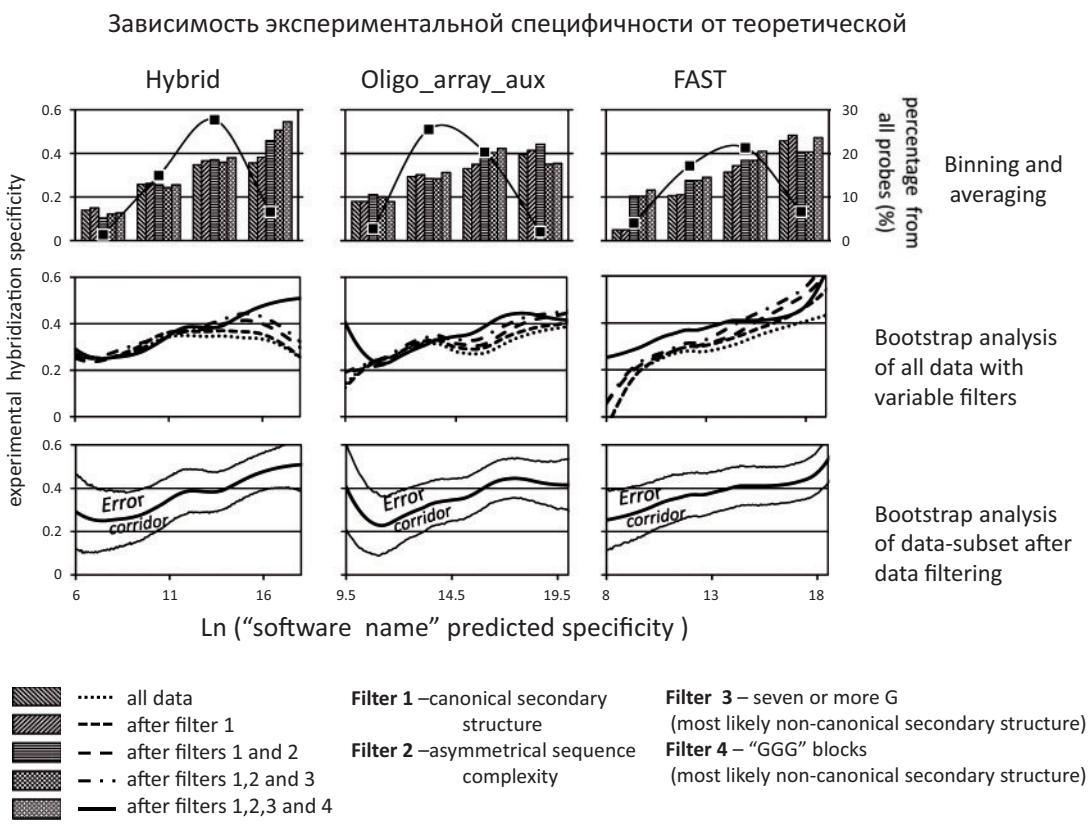

Fig. 3. Relationships between experimental and predicted hybridization specificities. Experimental hybridization specificity is shown in relation to predicted hybridization specificity. Outputs are shown for three different software variants after binning and averaging and bootstrap analysis with 10 000 resamplings for each bin. Top plots show results of the binning and averaging approach. Middle and bottom plots show results of bootstrap analysis. Middle plots show that filtering out the probes with 'negative characteristics' improves the relationship between experimental and theoretical hybridization specificity values. Bottom plots show the relationships between theoretical and experimental hybridization specificities after the probes with 'negative characteristics' were removed

significant stable differences were found between the most specific hybridization probes and the majority of remaining probes.

3.4 Suggested algorithm for selection of most specific oligo-probes

Based on the work described in the previous sections, we suggest the following steps for design of the most specific oligonucleotides for Affymetrix based platforms.

1. Filtering out probes with secondary structure (we suggest $\Delta G \leq -2$ kcal/mol).
2. Filtering out G rich probes (we suggest using probes with G counts of less than 8 nucleotides).
3. Filtering out the probes with at least one GGG motif.
4. Filtering out probes with low SC score (we suggest an SC score of 0.95 or higher).
5. Filtering out the probes with low oligo-target duplex stability, we suggest removing probes with less stability than $\Delta G \geq -26$ kcal/mol.
6. Filtering out probes with relaxed oligo-target duplex stability or with low predicted hybridization specificity, where thresholds for FASTH ln(S) = 14, for 'NCBI-Hybrid' ln(S) = 12, for OligoArrayAux ln(S) = 15, for Osaka University Software ln(S) = 32.

The suggested thresholds allow assigning approximately 10–20% of probes from Affymetrix tiling arrays to be specific with averaged specificity values above 0.45. The thresholds for filtering out probes with 'negative characteristics' were chosen iteratively by trial and error, these values could be further optimized according to new experimentally verified data. Both the final proportions of the remaining probes after five steps of filtering and average hybridization specificity of chosen probes, are threshold-dependent. If the goal of the optimization procedure is to identify a selection of several highly efficient probes for a specific gene, then more stringent thresholds could be applied (e.g. when the probe specificity is an order of magnitude higher than the average level of oligos in the input sequence).

4 Discussion

In our study we identified some 'negative characteristics' that could help in setting apart probes with particularly low hybridization specificity. The majority of these characteristics are related to the probes' physical features, which slow down both specific-hybridization and cross-hybridization in one way or another. For example, probes' canonical secondary structure formation allows probe self-interaction through Watson–Crick base pairing and diminishes their ability to interact with targets. This intra-molecular self-interaction is making probes less accessible to inter-molecular interaction and consequently slowing down the oligo-target duplex formation. The same is true for non-canonical interactions and secondary structure formations. It was reported earlier that probes with several G nucleotides in a row can be involved in Hoogsteen-Hydrogen pairing (Binder *et al.*, 2009; Fasold *et al.*, 2010; Langdon *et al.*, 2008; Memon *et al.*, 2010; Upton

et al., 2008; Wu *et al.*, 2007). We can also explain the negative influence of high G content or GGG blocks upon probes' hybridization specificity by a higher chance of the probes' self-interaction through Hoogsteen-Hydrogen pairing.

The theoretical specificity of hybridization was calculated in our study by using basic rules of equilibrium thermodynamics. Perhaps only a subset of target-interacting probes can achieve this equilibrium during array hybridization experiments. It is likely that our study allowed us to detect this subset by eliminating the probes that are involved in canonical or non-canonical interactions. Thus, the exclusion of such probes from the system allows better categorization of remaining probes according to theoretically calculated values.

Our work demonstrates that the specific hybridization signals in most CGH-array experiments are largely masked by cross-hybridization. This masking effect could be substantially diminished by improving the oligo-probe selection procedure during array design. The selection procedures described in this study can help detect probes with a ratio at least two fold greater than the majority of probes used for hybridization experiments today, where the ratio between specific- and cross-hybridization is between 25 and 75%.

Thus, additional pre-filtering of oligonucleotides by removing probes that are capable of forming canonical or non-canonical secondary structures or probes with low sequence complexity altogether improves this categorization. Our approach creates an efficient categorization of most specific versus least specific probes based on an estimation of duplex stability. Selection and calculation of duplex free energy between oligo-probes and partially complemented targets can also be applied for modeling and predicting probes' hybridization specificity (Supplementary Materials). However, taking into account that models with estimation of partially complemented targets are comparable in efficiency with the simplified approaches based only on perfect duplex stability, oligo-target duplex stability calculation is preferable due to its transparency and time efficiency. Perfect duplex stability between an oligo-probe and fully complemented target can be a strong discriminating factor between highly specific and non-specific hybridization probes.

In summary, the highest hybridization specificity can be attributed to probes that interact with their targets quickly and are not involved in parallel hybridization interactions. Such probes have the following sequence characteristics: high oligo-target duplex stability, low secondary structure stability, absence of GGG motifs, low G count and high sequence symmetry and complexity.

Funding

This work was supported by the Intramural Research Programs of the National Library of Medicine, National Institutes of Health [to A.Y.O. and S.A.S.]; by Russian Foundation for Basic Research [grant number 15-07-05783 to N.N.N.].

Conflict of Interest: none declared.

References

Binder,H. *et al.* (2005) Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. *Langmuir*, **21**, 9287–9302.

Binder,H. *et al.* (2009) Mismatch and G-stack modulated probe signals on SNP microarrays. *PLoS ONE*, **4**, e7862.

Chudin,E. *et al.* (2002) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.*, **3**, RESEARCH0005.

Fasold,M. *et al.* (2010) G-stack modulated probe intensities on expression arrays – sequence corrections and signal calibration. *BMC Bioinformatics*, **11**, 207.

Furusawa,C. *et al.* (2009) Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays. *Bioinformatics*, **25**, 36–41.

Hooyberghs,J. *et al.* (2009) The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. *Nucleic Acids Res.*, **37**, e53.

Kondrashov,A. and Shabalina,S. (2002) Classification of common conserved sequences in mammalian intergenic regions. *Hum. Mol. Genet.*, **11**, 669–674.

Landau,L. and Lifshitz,E. (1980) *Statistical Physics*, Part 1 (*Course of Theoretical Physics*, Volume 5). Elsevier Ltd., ISBN: 978-07506-3372-7.

Langdon,W. *et al.* (2008) Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips. *Brief. Bioinf.*, **10**, 259–277.

Matveeva,O. and Shabalina,S. (1993) Intermolecular mRNA-rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Res.*, **21**, 1007–1011.

Matveeva,O. *et al* (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Res.*, **31**, 4211–4217.

Matveeva,O. *et al.* (2007) Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Nucleic Acids Res.*, **35**, e63.

McCall,M. and Irizarry,R. (2008) Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Res.*, **36**, e108.

Memon,F. *et al.* (2010) A comparative study of the impact of G-stack probes on various Affymetrix GeneChips of mammalia. *J. Nucleic Acids*, **2010**, 489736.

Nechipurenko,Y. (2015) *Analysis of Binding of Biologically Active Compounds to Nucleic Acids*. ICI, Moscow-Izhevsk, p. 188.

Ogurtsov,A. *et al.* (2006) Analysis of internal loops within the RNA secondary structure in almost quadratic time. *Bioinformatics*, **22**, 1317–1324.

Ogurtsov,A. *et al.* (2008) Expression patterns of protein kinases correlate with gene architecture and evolutionary rates. *PLoS One*, **3**, e3599.

Ragan,C. *et al.* (2009) Transcriptome-wide prediction of miRNA targets in human and mouse using FASTH. *PLoS ONE*, **4**, e5745.

Rouillard,J.M. *et al.* (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.

SantaLucia,J. Jr. *et al.* (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.

Segal,E. and Widom,J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.*, **7**, 443–456.

Shabalina (2002) Region of intermolecular complementarity in Escherichia coli 16S rRNA, mRNA, and tRNA molecules. *Mol. Biol.*, **36**, 460–465.

Shabalina,S. *et al.* (2006) Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics*, **7**, 65.

Upton,G. *et al.* (2008) G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genomics*, **9**, 613.

Webb,C. *et al.* (2002) Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res.*, **30**, 1233–1239.

Weckx,S. *et al.* (2007) Thermodynamic behavior of short oligonucleotides in microarray hybridizations can be described using Gibbs free energy in a nearest-neighbor model. *J. Phys. Chem. B*, **111**, 13583–13590.

Wei,H. *et al.* (2008) A study of the relationships between oligonucleotide properties and hybridization signal intensities from NimbleGen microarray datasets. *Nucleic Acids Res.*, **36**, 2926–2938.

Wu,C. *et al.* (2007) Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics*, **23**, 2566–2572.

Xia,X.Q. *et al.* (2010) Evaluating oligonucleotide properties for DNA microarray probe design. *Nucleic Acids Res.*, **38**, e121.

Zhang,L. *et al.* (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.

Zhang,L. *et al.* (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.*, **35**, e18.