Model based probe set optimization for high-performance microarrays

Germán Gastón Leparc[*] Thomas Tüchler[*]
Gerald Striedner[†] Karl Bayer[†] Peter Sykacek[*]
Ivo L. Hofacker[‡] David P. Kreil[*§]

Revision 30th October 2008

— Supplement —

Enquiries: nar0807@kreil.org

[*] Chair of Bioinformatics,
Boku University, Vienna,
A-1190 Muthgasse 18, Austria
[†] Institute of Applied Microbiology,
Boku University, Vienna,
A-1190 Muthgasse 18, Austria
[‡] Theoretical Biochemistry Group, Institute for Theoretical Chemistry,
University of Vienna,
A-1090 Währingerstrasse 17, Austria
[§] Corresponding author

Required Declarations

The supplementary material which this document references is for confidential preview only. Please treat all data accordingly.

We cannot take responsibility for the availability of material referenced by links to external sites. Please advise us if you encounter any problems.

Main paper abstract

A major challenge in microarray design is the selection of highly specific oligonucleotide probes for all targeted genes of interest while maintaining thermodynamic uniformity at the hybridization temperature. We introduce a novel microarray design framework (TherMODO) that for the first time incorporates a number of advanced modelling features: 1) A model of position-dependent labelling effects that is quantitatively derived from experiment. 2) Multi-state thermodynamic hybridization models of probe binding behaviour, including potential cross-hybridization reactions. 3) A fast *calibrated* sequence-similarity based heuristic for cross-hybridization prediction supporting large-scale designs. 4) A novel compound score formulation for the integrated assessment of multiple probe design objectives. In contrast to a greedy search for probes meeting parameter thresholds, this approach permits an optimization at the probe set level and facilitates the selection of highly specific probe candidates while maintaining probe set uniformity. 5) Lastly, a flexible target grouping structure allows easy adaptation of the pipeline to a variety of microarray application scenarios. The algorithm and features are discussed and demonstrated on actual design runs.

This online supplement is archived at http://bioinf.boku.ac.at/pub/thermodo2008/. Data and scripts may also available from that archive. Software is available on request from the authors. (We can either run our code on your data, or collaborate in installing our system at your site.)

Contents

Chapter S-1

Using this archive

S-1.1 Viewing the Supplement and material referenced

This document is provided in PDF format (*cf.* Section S-1.2). Auxiliary information is referenced by HTTP URLs (Hyper Text Transfer Protocol – Universal Resource Locations). If you view this document in a stand-alone browser, *e. g.*, Acrobat Reader, clicking on a link should open a new browser window showing the content to which the link refers.

If you are viewing this document through a plug-in, your browser may loose the original page context when following a link, so when you go back to this document, you might return to the title page. In such a case you may want to save this document to a local disk, and then view it in a stand-alone PDF browser, like Acrobat Reader.

S-1.2 Description of file formats

File formats used in this archive include the following:

- American Standard Code for Information Interchange (ASCII) is used in data files, pre-formatted text for reports, and program code / script files. Columns in data files are typically Tab delimited.
  This format is the simplest and should cause the least problems. In particular, Tab delimted files are well viewed in any spreadsheet program.

- Adobe Portable Document Format (PDF) for typeset material. This supplement is made available in PDF.

  There are free viewer programs available for this format. To obtain such a viewer, please visit, for example:

  - Ghostscript, Ghostview and GSview from the Computer Sciences Department at the University of Wisconsin-Madison, USA,
  - Adobe Acrobat Reader from Adobe Inc., USA.

  Many browser programs for the World Wide Web can run so-called plug-ins for viewing PDF content.

- Adobe PostScript (PS) for typeset material. To obtain free tools for viewing and printing, please visit, for example, Ghostscript, Ghostview and GSview from the Computer Sciences Department at the University of Wisconsin-Madison, USA. These files are provided for convenience only, and are usually the best format for printing.

- `bzip2` compressed files. Large files (particularly text) may be compressed with `bzip2` for efficiency. Free utilities to unpack such files are available from `http://www.bzip.org/`.

- Various graphics file formats. Typical formats include JPEG, which is a lossy compression format well suited for photos with smooth gradients, and TIFF, which is a particularly flexible format, supporting both lossy and non-lossy compression schemes (TIFF-FAQ). For viewing or converting many graphics file formats, free tools are available (GraphicsMagick, ImageMagick).

- ZIP archives. Larger collections of files are provided in compressed archives. Free utilities to unpack these archives are available from the Info-ZIP group. Users of the Microsoft Windows system may wish to use WinZIP.

Chapter S-2

Methods

S-2.1 Implementation

The framework has been implemented as a series of interacting Perl scripts that make heavy use of nested data-structures and associative arrays (hashes/maps) for configuration, lookup, and caching. For memory efficiency, simple arrays are kept as packed structures. Parametrization and interaction of tools uses poor man's relational tables in form of flat text files. While not as robust as an RDBMS, this makes intermediary results easy to access from any computing environment.

S-2.2 Quantitative labelling model

TherMODO employs a quantitative model to calculate the probability of the labelling process generating a labelled product that includes the probe binding site. To demonstrate the impact of positional effects we selected a popular microarray protocol that was appropriate for the *E. coli* design discussed in the Manuscript: randomly primed labelling by reverse transcriptase incorporating amino-allyl dUTP (see S-2.2.2). In this section we motivate the model used and detail the experiments allowing parameter estimation. The model fit is discussed in the Results section (S-3.1).

S-2.2.1 Model construction

The probability that a particular template region becomes part of a labelled sequence is affected by enzyme processivity and primer type. We begin by introducing simple schematic models which serve as building blocks in the construction of more comprehensive, realistic models. Schematic illustrations of the models and their probability distributions are compiled in Figs S-2.1 and S-2.2.

Let us first consider models for full-length labelling (Fig. S-2.1, left-hand panels).

Full length, end-primed labelling

End-primed labelling with complete enzyme read through constitutes the most simple model, in the sense that all labelled products will be identical. Nucleotides at all positions x will be part of the labelled product with equal probability

$$P_L(x) = 1/n \ ,$$

for a template of length n (Fig. S-2.1, top left-hand panel). The model is an appropriate approximation for short transcript lengths and high-processivity labelling reactions. It can, *e. g.*, be applied to an oligo-(dT)-primed reverse transcriptase incorporating amino-allyl dUTP and is directly supported by the TherMODO pipeline.

Full length, random-primed labelling

Random-primed labelling with complete enzyme read through generates a set of labelled products starting at random template positions and all ending at the 5′-end of the template. Regions close to the 5′-end are hence more likely

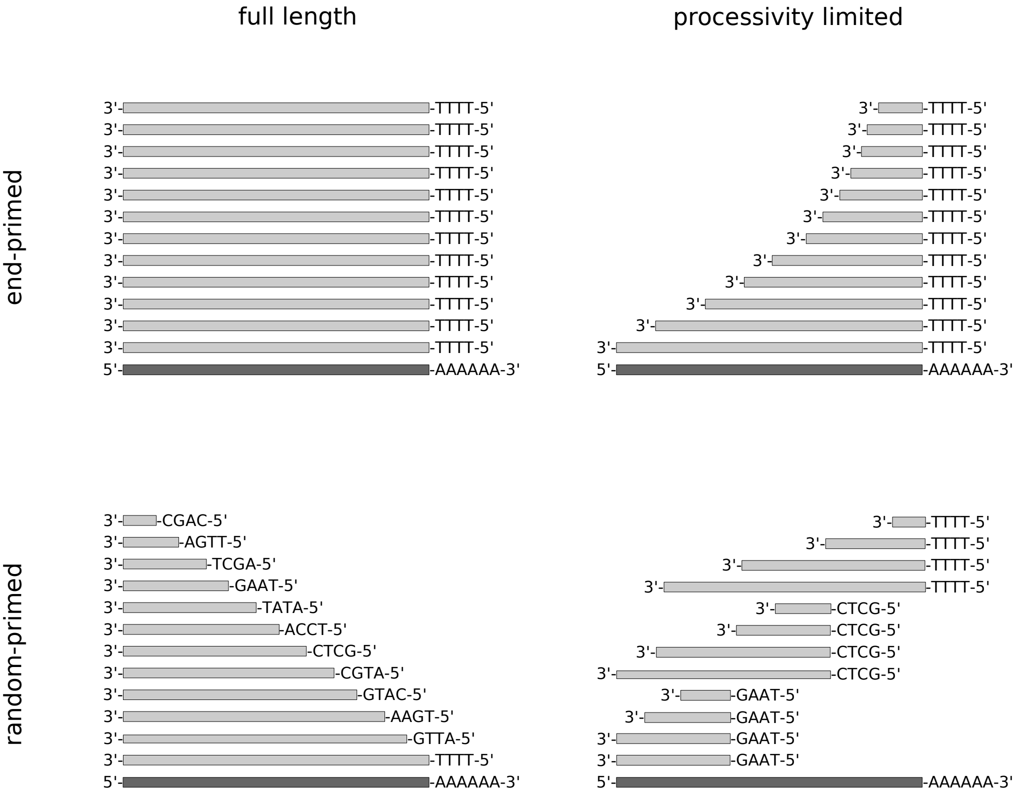


Figure S-2.1: Schematic illustration of different labelling models, with dark bars depicting templates and light bars representing labelled products. The probability that a region on the template is part of a labelled product is affected by the enzyme processivity and the type of primer used. Full-length labelling models in the left-hand panel are compared to enzyme processivity limited labelling in the right-hand panels. Top panels illustrate models for end-primed labelling, whereas bottom panels depict random-primed labelling.

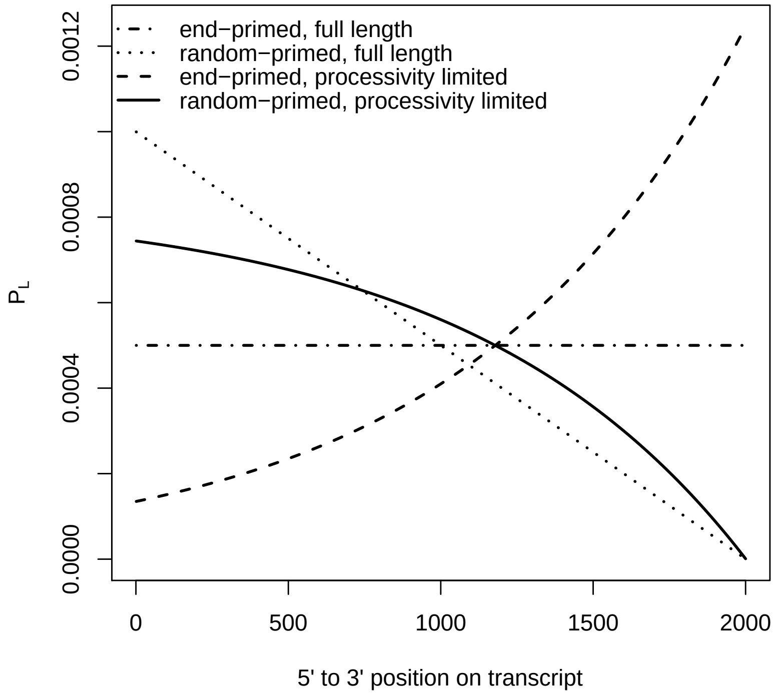


Figure S-2.2: The probability that a region on the template is part of a labelled product is affected by the enzyme processivity and the type of primer used and is shown on the y-axis. As an example, the pronounced positional labelling effects are illustrated for a 2000 bp long transcript and a characteristic length of 900 bp for each of the four models discussed.

because they are part of most products, short or long (Fig. S-2.1, bottom left-hand panel). The probability that a nucleotide will be part of a labelled product,

$$P_L(x) = x/Z \ ,$$

is thus proportional to the distance x from the $3'$-end of the template. The constraint $\sum_1^n P_L(x) = 1$ yields the normalization constant

$$Z = n\,(n+1)/2 \ .$$

The model is an appropriate approximation for short transcript lengths and high-processivity labelling reactions. It can, *e. g.*, be applied to a random-primed reverse transcriptase incorporating amino-allyl dUTP and is directly supported by the TherMODO pipeline.

Processivity limited, end-primed labelling

The models considered so far assume full-length labelling. In most experimental settings, however, this assumption is not a good approximation and labelling is frequently interrupted before reaching the end of the template. It is known that the secondary structure and composition of a template affects labelling enzyme drop off (1), and competitive binding further reduces labelling processivity in complex mixtures.

Let us consider a model for end-primed labelling where the labelling enzyme equally likely drops off the template at any labelling step or else continues with the probability c. A nucleotide at the distance x from the $3'$-end of the template will then be part of a labelled product with the probability

$$P_L(x) \ \propto \ c^x \ =: \ e^{-x/\lambda} \ ,$$

showing an exponential decay with the characteristic length λ (Fig. S-2.1, top right-hand panel). The characteristic length λ describes how rapidly longer products become more unlikely: a product that is λ nt longer is about 63% less likely. Under the constraint $\sum_1^n P_L(x) = 1$ one obtains the normalization constant

$$Z = \frac{c\,(1-c^n)}{1-c}\ .$$

This distribution provides, *e. g.*, an appropriate model for an oligo-(dT)-primed reverse transcriptase incorporating Cy-dye conjugated nucleotides, and related methods. Protocol particulars like the enzyme and label types will affect the characteristic length λ of the process, which needs to be measured. The TherMODO code supports this model for user-supplied λ.

Processivity limited, random-primed labelling

Here, we finally combine the effects of limited enzyme read through and the different starting positions of random-primed labelling along the template (Fig. S-2.1, bottom right-hand panel).

A nucleotide at the distance x from the 3′-end of the template will be part of a labelled product with the probability

$$P_L(c)\ \propto\ \sum_{i=0}^{x-1} c^i\ \propto\ 1-c^x\ ,$$

where i sums over all primer binding sites from x to the template 3′-end. The decay of the exponential component can again be characterized by a length-scale λ, so that $c^x\ =:\ e^{-x/\lambda}$. The probability distribution

$$P_L(x) = (1-c^x)/Z\ , \tag{S-2.1}$$

has the normalization constant

$$Z = \sum_{x=1}^{n}\Big(1-c^x\Big)\ =\ n - \frac{c\,(1-c^n)}{1-c}\ .$$

This distribution provides, *e. g.*, an appropriate model for a random-primed reverse transcriptase incorporating amino-allyl dUTP, and related methods. Protocol particulars like the enzyme and label types will affect the characteristic length λ of the process, which needs to be measured. The TherMODO code supports this model for user-supplied λ.

We will next demonstrate how λ can be obtained experimentally.

S-2.2.2 Experimental assay of labelling characteristics

A quantitative labelling model accounting for both random priming and limited labelling enzyme processivity, Eq. Eq. (S-2.2.1)), was fitted to measurements of an actual labelling reaction. We employed a popular labelling protocol that was appropriate for *E. coli* transcripts, which lack poly-A tails. The labelling protocol uses a random-primed reverse transcriptase to incorporate amino-allyl dUTPs into cDNA transcripts. Avoiding bulky dye-conjugated nucleotides that impede transcription, fluorescent dyes can then be conjugated to the amino-allyl modified nucleotides in a separate step.

For the labelling experiment, RNA extracted from *E. coli* strain HMS174(DE3) containing a pET11a (GFPmut.3.1) plasmid vector (2) was reverse transcribed using the AffinityScript HC Reverse Transcriptase component of the FairPlay III Microarray Labelling Kit (Stratagene, Cat. No. 252012). Amino-allyl dUTPs and random hexamer primers (MWG Biotech AG) were employed according to the instruction manual provided with the kit. Reverse transcripts were then, however, *not* conjugated with fluorescent dyes because the bulky dyes would interfere with the subsequent measurements of length distributions. The unlabelled RNA and the produced labelled cDNA samples were then analysed by capillary electrophoresis (Agilent 2100 Bioanalyzer, RNA nano LabChip) and on a 1% agarose gel (*cf.* Fig. S-3.1). Gel images were quantified using `imageJ` (http://rsb.info.nih.gov/ij/). With the help of the size markers (RiboRuler Ladder, Fermentas, cat #SM1812), the measured *fluorescence signal distributions* of the samples where then transformed into *molecule length distributions* (Fig. S-3.1).

S-2.2.3 Fitting the labelling model, simulations

Forward simulations were used to generate model predictions, iterating the below algorithm:

- Assume a template RNA molecule according to the observed distribution of RNA molecule lengths.
- Pick a random hexamer primer binding site along this RNA template.
- Reverse transcription.
    - Add nucleotides to the cDNA until the reverse transcriptase either reaches the end of the template RNA or until it drops off the template by random chance with probability c.
- Record the length of the resulting cDNA.

At convergence, this yields the distribution of the cDNA molecule lengths derived from simulated labelling reactions. Simulated and measured distributions could then be compared. Selecting the best simulation directly gives an estimate of the process parameters, in this case, the characteristic transcription length λ of Eq. Eq. (S-2.2.1) – see Results (S-3.1).

S-2.3 Probe binding site accessibility

For reasons of computational efficiency when considering many probes of different sizes for longer transcripts, we use two steps to calculate the probability P_A of a probe binding site being accessible, *i. e.*, not part of a stable secondary structure of the transcript at T_{hyb}. Exploratory studies of probe binding behaviour suggest that a stretch of 13–15 matching nucleotides (nt) can already give rise to detectable cross-hybridization (3), and it can thus be expected that regions of this size are typical seeding regions for duplex formation. In a first step we therefore use `RNAplfold` to obtain the accessibility of 13 nt regions in the transcript, where we set the size u of the unpaired region to 13 and the maximal span L and window size W both equal to 100.

The results for the short regions are then combined to calculate the probability that the entire probe binding site is accessible. For an oligo binding region of length l, the maximum number m of 13-mers that can be accommodated in this region is equal to $l/13$, while the number of offset spacings is the remainder f. The probability of accessibility starting at target position x is then calculated as

$$P_A = \frac{1}{f+1} \sum_{i=0}^{f} \prod_{j=1}^{m} P_{13}\Big(x + 13\,j + i - 1\Big) ,$$

from the probabilities P_{13} of 13-mers at particular positions in the target sequence being unpaired.

We have verified that this two-step procedure is a very good approximation of `RNAplfold` results for the entire binding site (data not shown).

The cross-match accessibilities are calculated similarly.

S-2.4 Calibrated heuristic – default model scores

Probe candidates without any associated cross-matches are assigned a 'default' integrated Cross-Match score that is calculated from a conservative estimate of P_B (regression minus two standard deviations, see Figs S-3.3(a) and S-3.3(b)) as well as the respective observed mean of P_L and P_A as an unbiased estimate. This correctly avoids 'infinitely' good probes when no sequence similarity hits to potential binding partners other than the target transcript were found.

With this approach, the cross-hybridization potential of probe candidates with no sequence similarity to cross-matches can be assessed in a manner consistent with how probes with identified cross-matches are treated, thus allowing a quantitative comparison of all probes.

S-2.5 Characterization of alternate probe designs

For a characterization of alternate probe designs we employed three popular publicly available microarray probe design tools. OligoRankPick (4), OligoArray 2.1 (5), and YODA (6) were each used to generate probes for a test set of 4,471 *E. coli* targets. Extending published designs (4, 6) for protein coding subsets of 4,289 and 4,237 sequences (YODA and OligoRankPick), the full target set includes non-coding RNAs which can be challenging due to their strong secondary structures. All programs were run with the default settings for their parameters, unless noted otherwise.

OligoArray allows for variable oligonucleotide lengths for improved probe uniformity. Probes were therefore designed with minimum and maximum probe lengths of 65 and 69 bp, which provides a good compromise of sensitivity and specificity (7).

Earlier microarray designs (4, 6) for *E. coli* have similarly used probe lengths of 60 (YODA) and 70 (OligoRankPick). As these tools design probes of fixed lengths, designs were run for a probe length of 65 bp.

OligoArray default temperature thresholds were adjusted for probe length to accept probe–target melting temperatures of 91–96 ℃, and reject probe secondary structure and cross-hybridization stable at 69 ℃ (8).

TherMODO probe design is detailed in the Manuscript.

All probe sets were characterized by probe binding strength (P_B), probe self-folding (P_P), target region accessibility (P_A), and positional labelling effects (P_L). See Manuscript for details on these calculations.

Chapter S-3

Results

S-3.1 Quantitative labelling model fit

We have demonstrated how the characteristic product length λ of a labelling process can be obtained experimentally. This was shown for a popular labelling protocol, random-primed reverse transcriptase incorporating amino-allyl dUTPs into cDNA transcripts (see S-2.2.2). In particular, this labelling protocol is also appropriate for the discussed *E. coli* design run. Oligo-(dT) primed labelling cannot be used for prokaryotic mRNAs, which are not poly-adenylated. The length distributions of the unlabelled RNA templates and labelled cDNA products were measured with an Agilent 2100 Bioanalyzer and on a 1% agarose gel. Compared to the RNA lengths, the length distribution of the corresponding cDNAs were shifted towards smaller molecule lengths, reflecting the effects of both limited labelling enzyme processivity and random priming (Fig. S-3.1, right-hand panels).

A model that accommodates both the effects of random priming and limited labelling enzyme processivity (Eq. Eq. (S-2.2.1)) was fit to the data. The length distribution of labelled cDNA products can then be predicted from the length distribution of mRNAs by forward simulation. Simulation results (dashed lines) could closely reproduce the observed average length distribu-

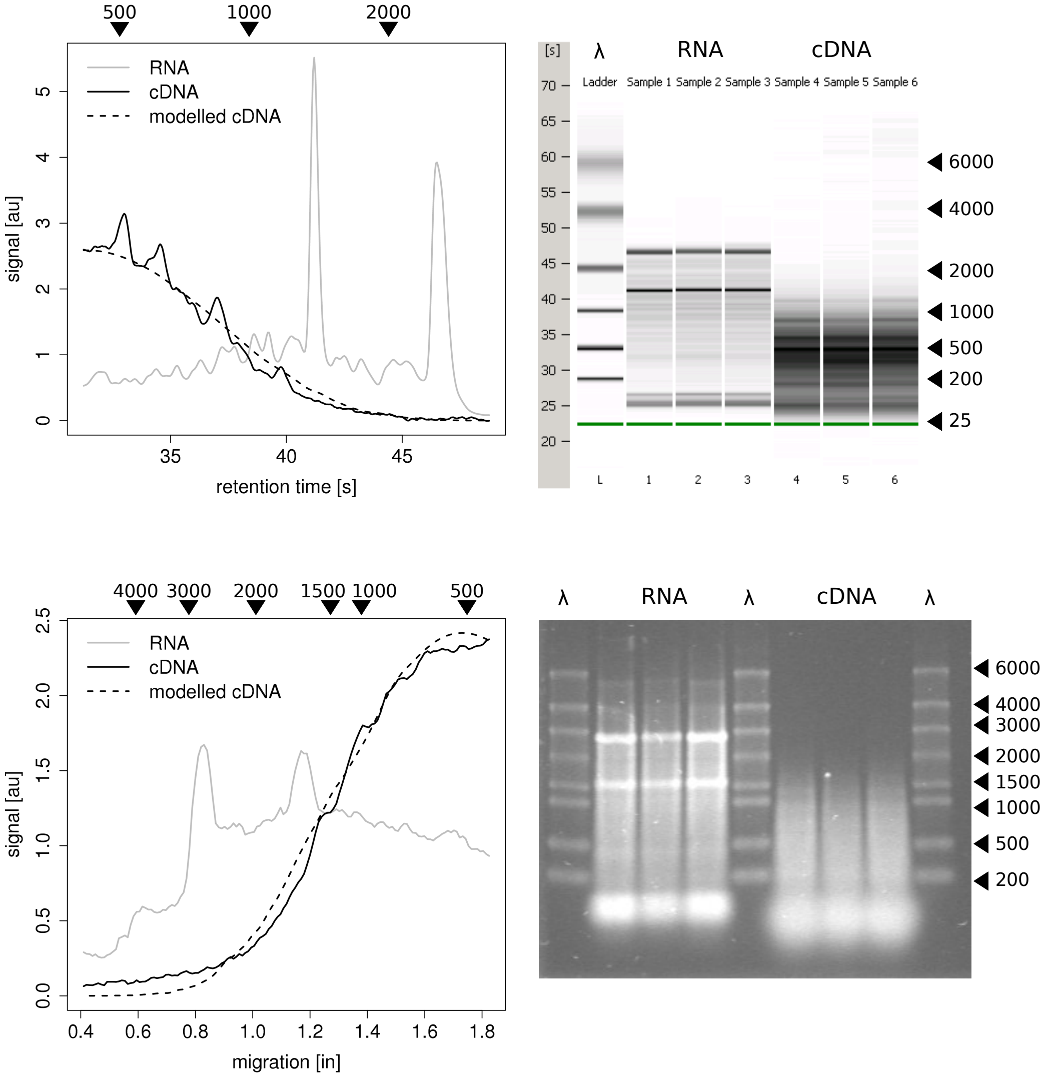
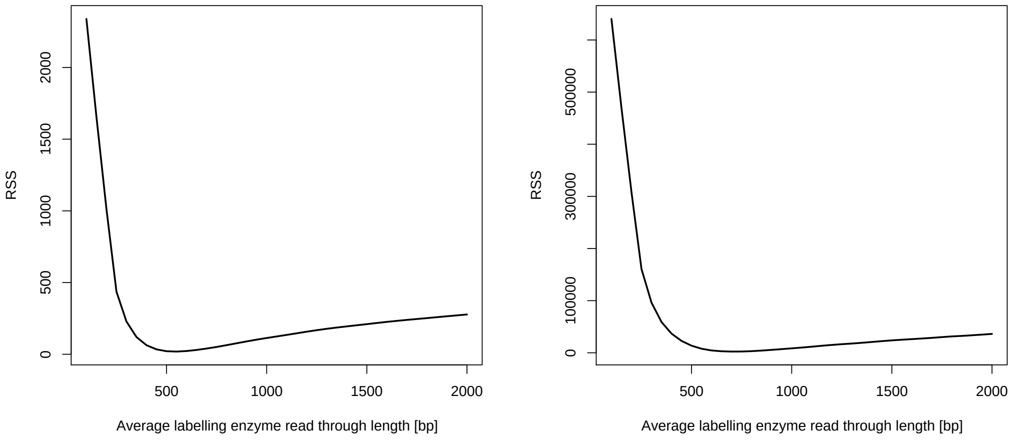


Figure S-3.1: Models simulating the reverse transcription reaction consistently reproduce the observed average length distribution of labelled cDNA products. This is shown for two complementary measurement methods. Length distributions of RNA and cDNA nucleotides were recorded by capillary electrophoresis (Agilent Bioanalyzer, top panels) and assessed on a 1% agarose gel (bottom panels). Arrow heads indicate ladder size markers. Bioanalyzer and gel RNA measurements show the typical ribosomal RNA peaks above 1000 bp as well as a pronounced tRNA peak below 200 bp. These do not, however, affect the model fit (data not shown). Model predictions (left-hand side panels, dashed lines) are well matched to the the average observed cDNA length distributions derived from either the Bioanalyzer (top) or the agarose gel measurements (bottom).



(a) (b)

Figure S-3.2: Comparison of the model fits for (a) Bioanalyzer data and (b) agarose gel measurements. The residual sum of squares (y-axis) is shown as function of the characteristic length λ (x-axis), with the lines plotting a Lowess average. Fits have been computed for $\lambda = 100, 150, 200, \ldots, 2000$. The plots show that, for the examined labelling protocol, a processivity with λ in the range 600–700 bp is supported by both independent assays.

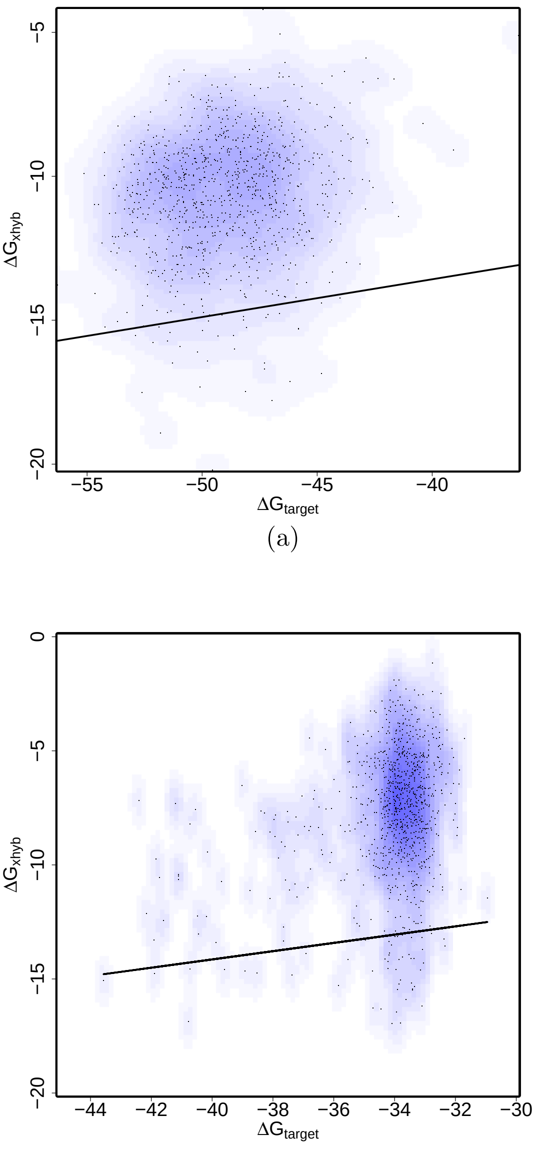
tions (solid lines) derived from either the Bioanalyzer data or the gel measurements (Fig. S-3.1, left-hand panels). Comparing model fits showed good agreement between the independent Bioanalyzer and gel assays (Fig. S-3.2). For the examined labelling protocol, both sets of measurements supported a processivity with a characteristic product length λ of about 600–700 bp. In the TherMODO probe design run for *E. coli* that is discussed in the Manuscript, positional labelling effects could thus be considered quantitatively using a typical characteristic length of 650 bp for the model introduced, Eq. Eq. (S-2.2.1).

Enzyme manufacturers report full length transcripts of several thousand basepairs when incorporating standard nucleotides with target-specific primers. It is noteworthy that the characteristic lengths obtained here are considerably lower. Several factors such as template secondary structure and sequence composition (1), nucleotide modifications, and a complex mixture of templates competing for reagents clearly affect the observed processivity of the enzyme under typical microarray labelling conditions. This explains the average positional effects (9) of probe binding sites along the target template and highlights the need for quantitative models of the labelling process already during microarray probe design.

S-3.2 Calibrated heuristic – regression

Sequence similarity based filters such as Blast are widely employed to manage the computational complexity of a comprehensive probe candidate search. Not only will these filters, however, miss relevant cross-matches compared to more sensitive thermodynamic calculations (6), the question more generally arises on how to quantitatively assess probe candidates that have some cross-hybridization potential *versus* probes with no Blast identified off-target sequence similarities ('cross-matches'). Current probe design methods assume that probes that have no identified off-target sequence identities are perfectly specific. As a result, probe candidates with undetected cross-hybridization potential are preferred over potentially better candidates with a small identified cross-hybridization potential, accepting probes with higher true cross-hybridization potential, stronger secondary structure, or labelling probability.

For both *E. coli* and human, random samples of 1000 probe candidates with no identified off-target sequence similarity were assessed for cross-hybridization potential by thermodynamic calculations. The `RNAduplex` tool was used to compute the probe–transcript binding strengths for all possible binding partners. Fig. S-3.3(a) displays results for *E. coli*, while Fig. S-3.3(b) shows the corresponding data for human. For each probe candidate, we plot the binding energy of the strongest cross-match. The relation between the probe–target binding energy ΔG_{target} and the worst cross-match binding strength ΔG_{xhyb} allows a conservative estimate by linear regression (black



(a)

(b)

Figure S-3.3: Calibration of the Blast heuristic prediction of cross-hybridization for probes with no cross-matches detected by sequence similarity. The x-axis plots the Gibbs free energy ΔG_{target} of probe–target binding and the y-axis shows the Gibbs free binding energy ΔG_{xhyb} of the strongest cross-match predicted by thermodynamic models. The black lines represent the conservative regression-based estimate. Panel (a) shows probes from *E. coli* and panel (b) shows human data.

line = regression trend minus two standard deviations). For *E. coli*, less than 4% of probes had cross-matches beyond the conservative estimate. Examining the respective regression based estimate for human, less than 5% of probes had cross-matches beyond the black line. Interestingly, one can very well use the regression based estimates interchangably (with only 1% of *E. coli* probes beyond the human based estimate and less than 6% of human probes beyond the *E. coli* based estimate).

One can thus make a conservative estimate of the cross-hybridization potential of probe candidates with no off-target sequence similarities, based on either organism, *e. g.*,

$$\Delta G_{\text{default}} = (0.13 \pm 0.02)\,\Delta G_{\text{target}} - (8.33 \pm 1.24)\;, \qquad (\textit{E. coli}\text{ regression})$$

as generic heuristic rule.

S-3.3 Characterization of alternate probe designs

The three alternate probe design programs examined represent different established approaches to probe design and have complementary features. For instance, YODA incorporates a custom sequence similarity search (`SeqMatch`) for the identification of potential cross-hybridization that is more sensitive than a BLAST run with typical parameters. OligoArray employs thermodynamic models for the assessment of probe–target duplexes, cross-hybridization, and self-folding. Both YODA and OligoArray use greedy search for selecting probes that match specified design criteria. In contrast, OligoRankPick chooses probes from a pool of candidates per target using a weighted rank-sum strategy for a number of probe qualities such as probe specificity GC-content, self-binding, and sequence complexity. With the exception of YODA, the tools employ a BLAST-based filter for identifying cross-hybridization.

For a comprehensive assessment of the compiled probe sets, full model ther-

Program & Link	Transcripts with probes	Unique probes	Unique corresp. probes TherMODO
TherMODO	4,471	4,381	–
OligoRankPick	4,471	4,357	4,381
OligoArray	4,222	4,157	4,166
YODA	4,110	4,110	4,110

Table S-3.1: Transcript coverage, overview. The number of transcripts covered by each design is shown. The number of unique probes limits the number of distinct transcripts that can be discriminated. The last column shows the number of unique TherMODO probes for the transcripts covered by each of the other designs. The program name links to a TAB-delimited text table of probe design results.

modynamic calculations were directly applied to all probes and their potential binding partners. The resulting measure of specificity ΔI is, in particular, independent of the algorithms and any sequence-similarity based heuristics employed in the original design processes.

Probe design results (follow the *links* in Table S-3.1) reflect the different design strategies and probe selection criteria of the examined tools. The achieved target coverage varied considerably between the examined tools. The number of transcripts for which probes could be designed is shown in Table S-3.1, together with the number of unique probe sequences constructed for each design, which reflects the maximum number of targets that can actually be descriminated. The OligoRankPick and TherMODO designs, both based on non-greedy probe selection, had the highest target coverage and featured higher numbers of unique probe sequences.

Figure S-3.4 provides a number of plots relating probe properties of the different designs. There are three blocks of panels, showing results for OligoRankPick (columns 1–3), OligoArray (columns 4–6), and YODA (columns 7–9). The columns in each block consider measures of binding strength P_B, signal intensity I_{tm} and specificity $\Delta I = I_{tm}/I_{xm}$. In the top row scatterplots, each dot represents the probe for a particular transcript, with the TherMODO value on the x-axis, and the y-axis showing the value for one of the other tools. The middle row compares the distributions of values between designs. Finally, the bottom panels display the distribution of differences per probe, with the dashed line showing the median. Instead of exploiting sequence-similarity based heuristics for speed, in this assessment, full model thermodynamic calculations were applied to all probes and their potential binding partners.

On average, the sensitivity of probes designed by TherMODO compares favourably, with a higher I_{tm} for about three in four transcripts, and typical probe sensitivity improving by 11–17% (see statistics and Fig. S-3.4, bottom row, middle column in each block). While all tools designed probes of excellent specificity ($\Delta I > \Delta I' = 10^{12}$) for the majority of genes, differences could be observed for more difficult subsets of about 6–12% of design targets. In particular, probes from the TherMODO design were more specific than probes from the other three designs, with a (median) 1000-fold improvement in probe specificity ΔI (see statistics and Fig. S-3.4, bottom row, third column in each block). Only a single TherMODO probe had slightly lower specificity than the corresponding OligoRankPick designed probe: This alternative probe, however, had an unusually low sensitivity. The slight reduction in specificity ($RT \log \Delta I = -1.5$), was thus more than made up by the improved sensitivity ($RT \log \Delta I_{tm} = +5.7$) while at the same time contributing to improved overall probe set unformity.

So as to also allow direct comparisons between the three alternate probe designs, we also examined probe properties for the subset of 4,049 targets that were common to all designs. Results remained similar (see statistics), with the non-greedy approaches (TherMODO and OligoRankPick) faring better than the other tools.

For the probe lengths studied here, performance typically improves with length (7). As some tools consider probes of varying lengths whereas others design probes of fixed lengths, it is interesting to separate length effects, *e. g.*, by focussing on a subset of 65-mer probes. Figure S-3.5 plots probe characteristics for 65-mers. The results remained similar with TherMODO having better overall sensitivity and specificity (see statistics).

In summary, through the above steps, we have shown by multiple criteria that probe characteristics of the TherMODO design compared favourably to designs by the examined established alternate tools. In particular, we could test the robustness of the observed probe qualities by subtracting probe-length effects and ensuring that the heuristic score did not unfairly skew results.

For a complete compilation of the data as well as the additional plots for the studies discussed in this text, please refer to the tables S-3.2 and S-3.3 below.

Oligonucleotide probe design characterization study: statistics and plots	
Full scores	
All pairwise comparisons between designs	Fig. S-3.4
All pairwise comparisons between designs, for 65-mers only	Fig. S-3.5
All comparisons for designed probes of targets common to all designs	Plots
All comparisons for designed probes of targets common to all designs, for 65-mers only	Plots
Raw scores	
All pairwise comparisons between designs	Plots
All pairwise comparisons between designs, for 65-mers only	Plots
All comparisons for designed probes of targets common to all designs	Plots
All comparisons for designed probes of targets common to all designs, for 65-mers only	Plots

Table S-3.2: Table of all statistical data and plots for the characteriziation of different probe designs. Results with Raw Scores are included to allow an examination of the effect of different thresholds $\Delta I'$. The results in the Full Scores section use the conservative $\Delta I' = 10^{12}$ motivated in the manuscript.

Probe design	TherMODO evaluation
TherMODO	Probe evaluation
OligoRankPick	Probe evaluation
OligoArray2	Probe evaluation
YODA	Probe evaluation

Table S-3.3: Table of different probe design sets and their corresponding characterization by TherMODO.

Bibliography

[1] Malboeuf,C.M., Isaacs,S.J., Tran,N.H. and Kim,B. (2001) Thermal effects on reverse transcription: improvement of accuracy and processivity in cdna synthesis. *Biotechniques*, **30**, 1074–8, 1080, 1082, passim.

[2] Reischer,H., Schotola,I., Striedner,G., Potschacher,F. and Bayer,K. (2004) Evaluation of the gfp signal and its aptitude for novel on-line monitoring strategies of recombinant fermentation processes. *J Biotechnol*, **108**, 115–125.

[3] 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**, 4552–7.

[4] Hu,G., Llinas,M., Li,J., Preiser,P.R. and Bozdech,Z. (2007) Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. *BMC Bioinformatics*, **8**, 350.

[5] Rouillard,J.M., Zuker,M. and Gulari,E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res*, **31**, 3057–62.

[6] Nordberg,E.K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–70.

[7] Chou,C.C., Chen,C.H., Lee,T.T. and Peck,K. (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res*, **32**, e99.

[8] Kreil,D.P. and Russell,R.R. (2005) There is no silver bullet–a guide to low-level data transforms and normalisation methods for microarray data. *Brief Bioinform*, **6**, 86–97.

[9] Kakuhata,R., Watanabe,M., Yamamoto,T., Obana,E., Yamazaki,N., Kataoka,M., Ooie,T., Baba,Y., Hori,T. and Shinohara,Y. (2008) Importance of probe location for quantitative comparison of signal intensities among genes in microarray analysis. *J Biochem Biophys Methods*, **70**, 926–931.

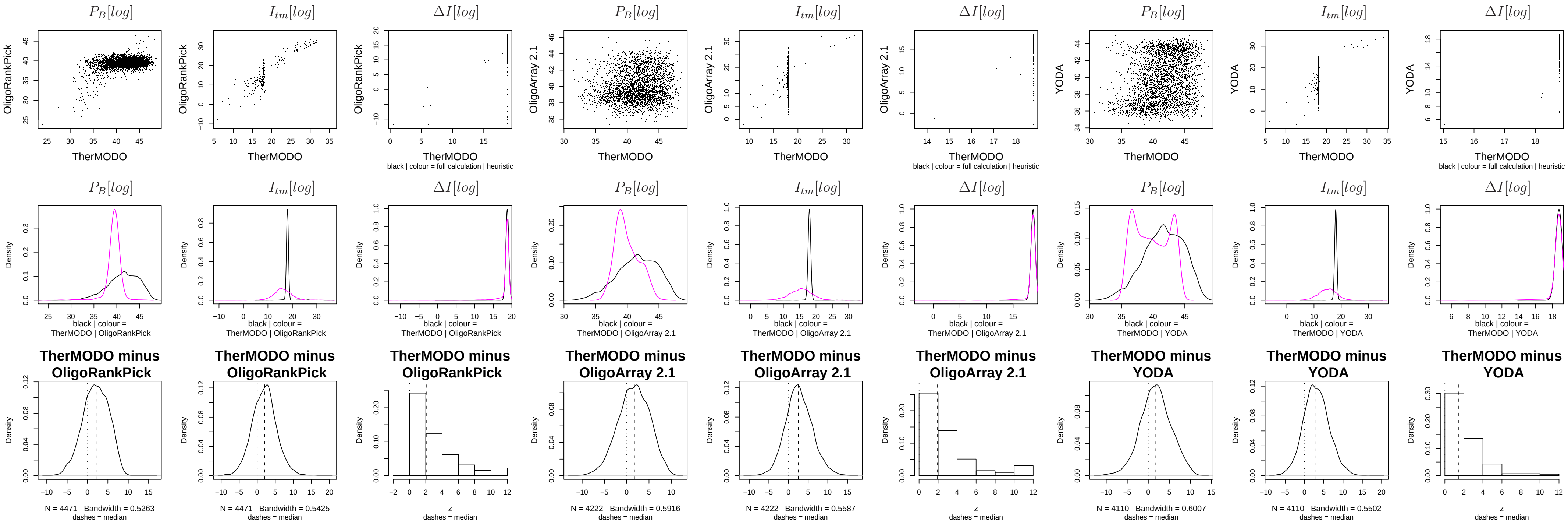


Figure S-3.4: Characterization of probe properties from *E. coli* design runs. Instead of exploiting sequence-similarity based heuristics for speed, in this assessment, full model thermodynamic calculations were applied to all probes and their potential binding partners. TherMODO probes are compared to probes designed by three popular tools: OligoRankPick (columns 1–3), OligoArray (columns 4–6), and YODA (columns 7–9). The columns in each set consider binding strength P_B, signal intensity I_{tm} and specificity $\Delta I = I_{tm}/I_{xm}$. In the top row scatterplots, each dot represents the probe for a particular transcript, the TherMODO value on the x-axis, and the y-axis showing the value for one of the other tools. The middle row compares the distributions of values. The bottom row plots the distribution of differences per probe, with the dashed line showing the median. Probes of perfect specificities in both designs did not contribute to the displayed histogram of ΔI differences. All values are shown on a $\log_{10}$ scale. For comparison, the same plots are shown in Fig. S-3.5 for subsets of probes with equal lengths 65.

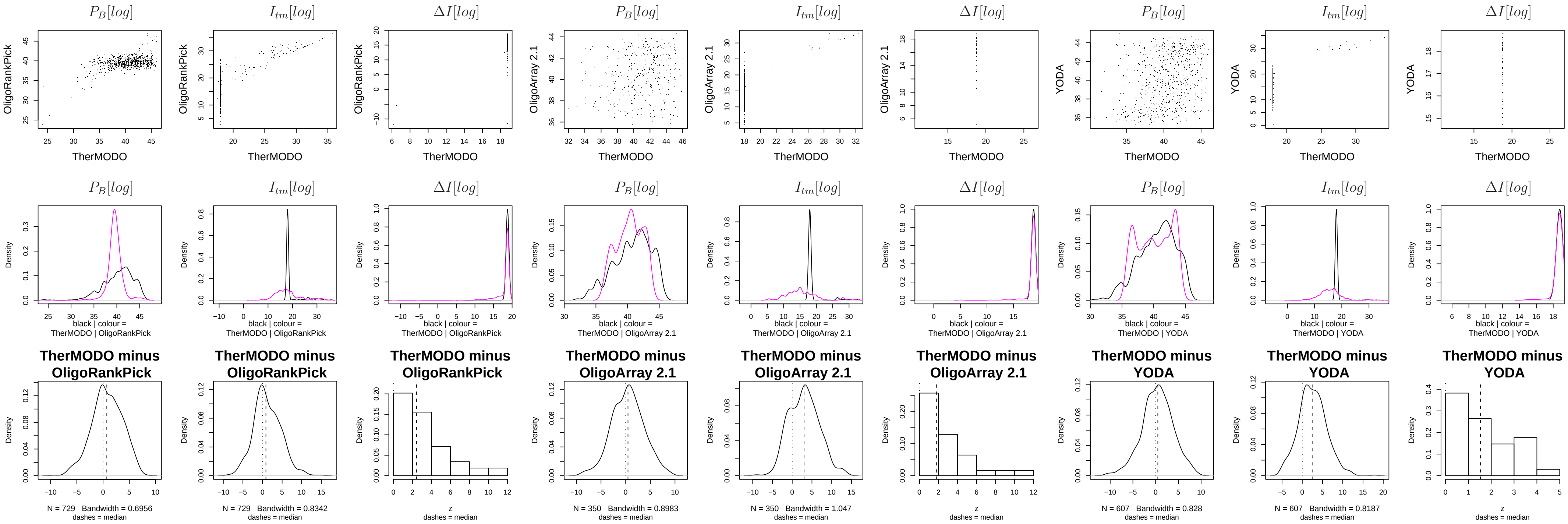


Figure S-3.5: Characterization of probe properties for 65-mers from *E. coli* design runs. Instead of exploiting sequence-similarity based heuristics for speed, in this assessment, full model thermodynamic calculations were applied to all probes and their potential binding partners. TherMODO probes are compared to probes designed by three popular tools: OligoRankPick (columns 1–3), OligoArray (columns 4–6), and YODA (columns 7–9). The columns in each set consider binding strength P_B, signal intensity I_{tm} and specificity $\Delta I = I_{tm}/I_{xm}$. In the top row scatterplots, each dot represents the probe for a particular transcript, the TherMODO value on the x-axis, and the y-axis showing the value for one of the other tools. The middle row compares the distributions of values. The bottom row plots the distribution of differences per probe, with the dashed line showing the median. Probes of perfect specificities in both designs did not contribute to the displayed histogram of ΔI differences. All values are shown on a $\log_{10}$ scale.

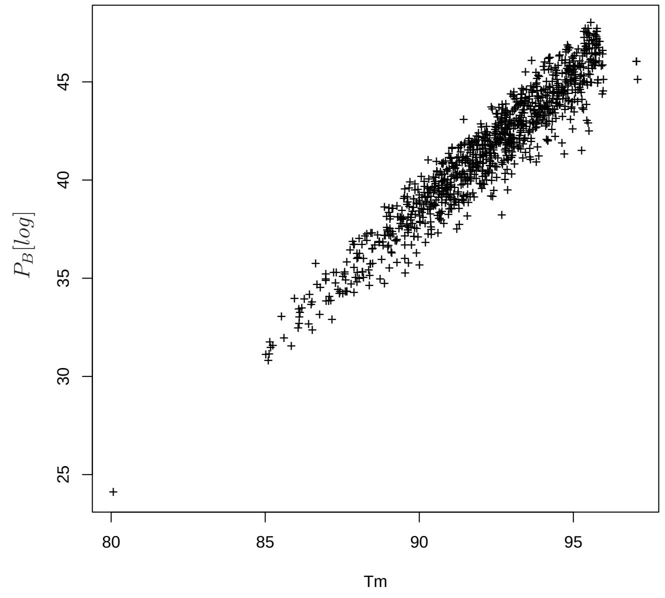


Figure S-3.6: Scatter plot of the melting temperature (T_m) on the x-axis and $P_B[log]$ on the y-axis for TherMODO probes. For illustrative purposes, the plot is generated from a random sample of 1000 probes.

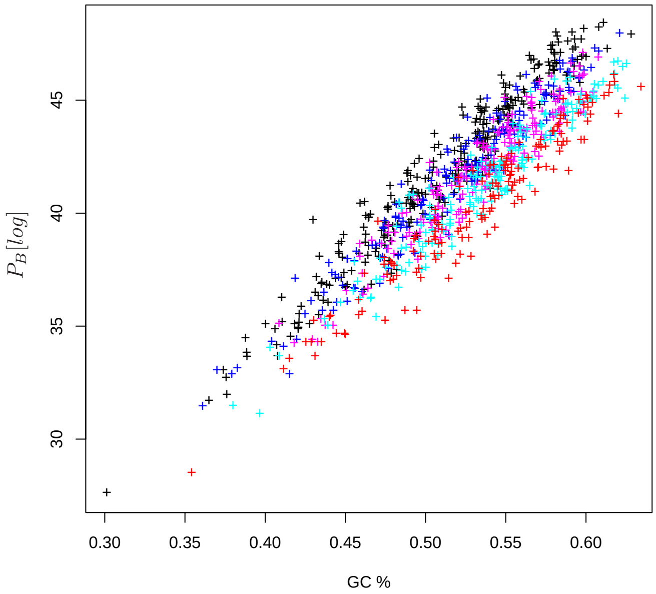


Figure S-3.7: Scatter plot of the GC content the x-axis and $P_B[log]$ on the y axis for TherMODO probes. For illustrative purposes, the plot is generated from a random sample of 1000 optimal probes with a little random jitter added to the GC content to improve display. Probes of length 69 are in black, while 68-mers, 67-mers, 66-mers, and 65-mers are in blue, magenta, cyan, and red, respectively.