Microarray Oligonucleotide Probes

David P. Kreil,[1] Roslin R. Russell,[2] and Steven Russell[2]

[1] Chair of Bioinformatics,
Department of Biotechnology,
Boku University Vienna
Muthgasse 18
A-1190 Vienna, Austria

[2] Department of Genetics,
University of Cambridge
Downing Street, Cambridge CB2 3EH, U.K.

Contact: MethEnz05@Kreil.Org
R.Russell@gen.cam.ac.uk
S.Russell@gen.cam.ac.uk

Supplement

Table and Figure numbers are continued from the main manuscript.

Practical considerations in probe-sequence design, a case study – technical issues

Construction of the target transcript set

In making a set of sequences non-redundant, the headers (names) of redundant sequences are usually merged, leading to very long sequence headers. This can at times trigger a malfunction in the BLAST program. The latest version of BLAST (Altschul *et al.*, 1997) was employed but the problem also affected earlier versions. The cause for this malfunction was not further investigated. The names of all target sequences hence had to be reduced to unique headers not longer than 60 characters.

Employment and post-processing

Minor changes to the OligoArray 2.1 source code allowed us to work around a disruptive BLAST output bug. This fix will be made available in the next version of OligoArray to be released later this year (J.-M. Rouillard, *pers.comm.*, 2005). We also adapted the sources to allow multiple instances of the program to share a working directory which simplified distributed dispatch.

Microarray production and hybridization protocols – effects on specificity

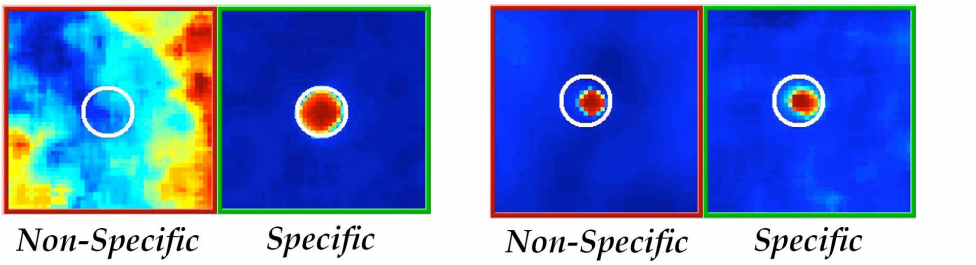

Figure 9. Influence of experimental conditions on hybridization specificity. False-colour images are shown of two-channel scans. In the sample, the target probed was present in only one of the channels. The other channel hence shows non-specific hybridization (left-hand side images). The first panel shows the results desired – no fluorescence detected in that channel. For the second panel, only the substrate chemistry and spotting buffer were changed. This was sufficient, however, for the same probe and sample to hybridize non-specifically, giving an artefact signal in both channels (Kreil *et al.*, unpublished data).

Discrimination of highly similar targets – the complexity of alternative splicing

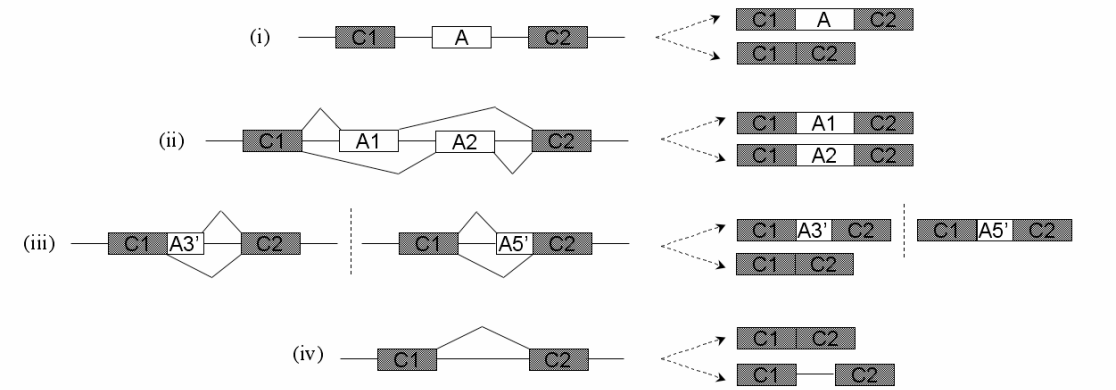

Figure 10: Alternative splicing. Exons and introns are represented by boxes and lines, respectively. There are four main types of alternative splicing: (i) *Single cassette exon inclusion/exclusion.* C1 and C2 are constitutive exons, *i.e.*, included in all splice forms, and flank a single alternative exon (A) that is included in one splice form and excluded in the other. (ii) *Mutually exclusive exons*. One of the two alternative exons A_1 and A_2 may be included in the splice form, but not both. (iii) *Alternative 3' (**donor**) and alternative 5' (**acceptor**) splicing sites.* Both exons are constitutive, but may contain alternative donor and/or acceptor splicing sites. (iv) *Intron inclusion*. An intron may be included in the mature mRNA strand. (Adapted from Shai *et al.*, 2004)

An overview of selected tools for microarray probe design

In a search for specific probes, non-specific hybridization is either predicted with more or less crude thermodynamic models alone, by sequence similarity and heuristics as substitute for these, or by a combination of the two. Tools exemplary for these approaches are:–

- PROBESEL uses dynamic programming supplemented by a suffix tree to maximize probe–target T_m while minimizing probe–non-target T_m (Kaderali and Schliep, 2002). Although the search is comprehensive and accounts for mismatches and bulges, more complex structural motifs like multi-loops are not considered to achieve reasonable running times. An optimal hybridization temperature is also calculated for the derived probe set.
- ProbeSelect obtains probe candidates with a maximal number of mismatches to non-targets using a suffix tree to map unique sequences, followed by either substring frequency based heuristics or a search based on dynamic programming (Li and Stormo, 2001). A heuristic approximation is used to estimate T_m for a set of heuristically selected non-target transcript regions that are expected to contribute to cross-hybridization.
- Oliz specifically targets probes to the 3'-UTR (untranslated region) of target transcripts (Chen and Sharp, 2002). Selection of probes for specificity is done exclusively by sequence similarity search (BLAST).

Table 2 lists a short selection of popular tools for probe design, compiled April 2006. We are aware that there are large numbers of additional programs available for probe design. If you wish to alert us to a particular tool that we do not list here, we look forward to hearing from you.

Subsequent tables provide a survey of tool characteristics and methods employed. Note that the survey tables present information primarily collected from scientific journals by one of the authors (RRR) through the course of a survey. For individual tools, more detailed information can often be obtained by examination of program sources, where available. If you can complete or correct information in these tables, we should be grateful to hear from you.

Predictions of probe–target melting temperature

Effective temperatures

It should be noted that all melting temperatures calculated by probe design tools are 'effective' temperatures. Hybridization is strongly affected by hybridization conditions, including buffer additives. This can already be seen in the formula relating thermodynamic quantities to the melting temperature,

$$T_m = \Delta H / \Delta S + R \ln (C / f)) + 12.0 \times \log_{10} [Na^+] - 273.15$$

where ΔH is the enthalpy; ΔS is the entropy change of the nucleation reaction and is based on the sequence content of an oligonucleotide and its target region using the updated N-N parameter tables (Allawi and SantaLucia, 1997); R is the universal molar gas constant, R = 8.314472(15) J/M/K = 1.9872 kcal/mol/K, with the thermodynamic calorie being 4.184 J; and f is 1 for self-folding and 4 for hybridization of different partners; C is the molar concentration of total oligonucleotides in the microarray experiment; and $[Na^+]$ is the molar concentration of sodium ions. Typically, however, since this is unknown in normal microarray experiments, $C = 10^{-6}$ M and $[Na^+]$ = 1 M

are used (Li and Stormo, 2001; Kaderali and Schliep, 2002; Rouillard *et al.*, 2003; Chou *et al.*, 2004). In addition, other salts like Mg^{++} and additives like formamide will also affect the melting temperature. In practice, as a full model of all these effects is usually not feasible, theoretical modelling hence yields effective temperatures that need to be empirically calibrated against physical temperature (*cf.* 'INDAC microarray probe design, validation and calibration experiments', Kreil *et al.*, in preparation).

The Nearest Neighbour (NN) Model

The most simple two-state models assume that DNA duplexes form like a 'zip'. The duplex is assumed to initiate at one end of the shorter of the two sequences and each base pair then forms sequentially from the initiation site. The model assumes that the total enthalpy and entropy of the duplex is the sum of the contribution of each neighbouring pair of base pairs in the duplex, thus taking into account base pairing energies and base-stacking energies. Thermodynamic parameters for the 10 possible correct pair-wise Nearest Neighbour (NN) interactions and various types of mismatch have been determined empirically (Breslauer *et al.*, 1986; SantaLucia *et al.*, 1996; Allawi and SantaLucia, 1997; Allawi and SantaLucia, 1998c; Allawi and SantaLucia, 1998b; Allawi and SantaLucia, 1998d; Allawi and SantaLucia, 1998a; Peyret *et al.*, 1999; Sugimoto *et al.*, 1996) and these can be used to calculate the total enthalpy, entropy, free energy and melting temperature for arbitrary sequences. After extensive independent verification and reviews, the established 'unified' free energy model parameters are now generally considered the most accurate and earlier approaches should not be applied.

Oligonucleotide selection tools based on NN model predictions either implement the necessary calculations internally or use external packages such as `melting` (Le Novere, 2001) or `mfold` (Zuker, 2003).

Empirical heuristics

For a very fast heuristic screen, oligonucleotide duplex melting temperatures are sometimes estimated from the GC content and length of a duplex. The simplest and crudest of them is the '**Wallace** rule' (Wallace *et al.*, 1979) which states the melting temperature (T_m) in degrees centigrade is composed of multiples of the sums of the incidence of A and T residues and G and C residues:

$$T_m = 2(A+T) + 4(G+C)$$

Another simple model, referred to as the '**Schildkraut'** calculation, takes into account GC composition as well as a dependence on duplex length and salt concentration:

$$T_m = (64.9 + 41) \times (gcCount / oligoLength) - (600 / oligoLength)$$

where *gcCount* is the number of all Gs and Cs in an oligonucleotide and, for microarrays, the molar concentration is often taken to be 0.1M (Schildkraut, 1965).

There also are other heuristic approaches for the calculation of melting temperature based on the GC composition of DNA/DNA duplexes, like the **Howley** model (Howley *et al.*, 1979), where

$$T_m = 81.5 - 16.6 \times \log (M/1 + 0.7M) + 41(XG+YC) - 500/L - 0.62\ F$$

Here *M* is the monovalent cation concentration (*XG+YC*), which represents the fractional percentages of the sequence that comprise cytidine and guanidine, *L* is the duplex length and *F* is the concentration of formamide.

Relative merit and ongoing developments

The accuracy of the predictions of these models have been tested in a study for a number of short (>25 bp) duplexes (Rychlik and Rhoads, 1989), which showed that the nearest neighbour model is the most accurate and robust algorithm for the analysis of short oligonucleotides. Our understanding of short oligonucleotides in solution can both be extrapolated to longer duplexes and more complex bound structures with loops, bulges, *etc*. as well as to the binding properties of tethered probes (see main text of review). It is noteworthy, however, that complex compound binding structures, for example involving regions of the same probe binding itself while other regions binding a target molecule are expected to be of greater relevance for longer probes and would need to be considered in accurate predictions.

Probe and target secondary structure

While the propensity of a probe to form a stable secondary structure is calculated by many probe design tools, to date, there are no tools that consider the secondary structure of the target. The fold-back of the target sequence onto itself and target–target interactions can impede probe-target binding, and consideration of such cases would require identifying secondary structures as part of assessing all probe–target and target–target interactions. Such calculations are computationally costly and would extend the oligonucleotide design time quite considerably, which may be the major reason why this has not been pursued by any of the current oligonucleotide selection tools, despite the impact of the effect being well studied (Koehler and Peyret, 2005).

Design Tool	Version	URL
ArrayOligoSelector		http://arrayoligosel.sourceforge.net/ (Bozdech *et al.*, 2003)
GoArrays		http://www.isima.fr/bioinfo/goarrays/ (Rimour *et al.*, 2005)
OligoArray	2.1	http://berry.engin.umich.edu/oligoarray2_1/ (Rouillard *et al.*, 2002; Rouillard *et al.*, 2003)
OliCheck		http://www.genomic.ch/techno_array.php (Charbonnier *et al.*, 2005)
Oligodb		http://oligodb.charite.de/ (Mrowka *et al.*, 2002)
OligoDesign		http://oligo.lnatools.com/expression/ (Tolstrup *et al.*, 2003)
OligoPicker		http://pga.mgh.harvard.edu/oligopicker/index.html (Wang and Seed, 2003)
OligoWiz	2.0	http://www.cbs.dtu.dk/services/OligoWiz/ (Nielsen *et al.*, 2003; Wernersson and Nielsen, 2005)
Oliz		http://www.utmem.edu/pharmacology/otherlinks/oliz.html (Chen and Sharp, 2002)
Osprey		http://osprey.ucalgary.ca/ (Gordon and Sensen, 2004)
Picky		http://www.complex.iastate.edu/download/Picky/tutorials.html (Chou *et al.*, 2004)
PRIMEGENS		http://compbio.ornl.gov/structure/primegens/ (Xu, 2000)
PROBESEL		http://www.zaik.uni-koeln.de/bioinformatik/arraydesign.html (Kaderali and Schliep, 2002)
ProbeSelect		Available on request, F. Li & G. Stormo, [lif, stormo]@ural.wustl.edu (Li and Stormo, 2001)
Promide		http://oligos.molgen.mpg.de/ (Rahmann, 2003)
ROSO		http://pbil.univ-lyon1.fr/roso/ (Reymond *et al.*, 2004)
YODA		http://pathport.vbi.vt.edu/YODA/ (Nordberg, 2005)

Table 2: A selection of popular probe design tools.

Design Tool	Sequence Similarity Search	Contiguous Identity	% Identity	Target / Probe Mismatch Pos.	Forward / Reverse Strand Match
ArrayOligoSelector	BLAST	No	?	No	No
GoArrays	BLAST, $w = 7$	Yes	Yes	No	?
OligoArray	BLAST, $w = 7$	?	No	No	No
OliCheck	BLAST	Yes	No	Yes	Yes
Oligodb	BLAST	No	No	No	No
OligoDesign	BLAST, $w = 9$	No	Yes	No	No
OligoPicker	BLAST, $w = 8$	Yes	Yes	No	No
OligoWiz	BLAST	Yes	Yes	No	?
Oliz	BLAST	Yes	Yes	No	No
Osprey	BLAST	Yes	?	No	Yes (?)
Picky	Suffix array	Yes	Yes	No	Yes
PRIMEGENS	BLAST	No	?	?	?
PROBESEL	Suffix tree	No	Yes	No	?
ProbeSelect	Suffix array	Yes	No	No	Yes
Promide	Suffix array	No	No	No	?
ROSO	BLAST, $w = 7$	No	Yes	No	?
YODA	SeqMatch, $w = 4$	Yes	Yes	No	?

Table 3: Features of Selected Freely Available Oligonucleotide Probe Design Tools (I). *Legend:* over.

Legend for Table 3 (alphabetic):

?: Not known from information published in scientific journals / unclear.

BLAST (Basic Local Alignment Search Tool) is a method for rapid searching of nucleotide and protein databases (Altschul *et al.*, 1990) by sequence similarity. Most oligonucleotide selection tools rely on BLAST, which uses a word-based look-up approach with a minimum word size of *w* nucleotides (nt) for DNA sequences. Note that sequences that have no common word of size *w* will be missed by such a search. Configurations employing a large word size for similarity searches in filters during oligo-design hence run a serious risk of failing to detect potential cross-hybridizations with consequently reduced oligonucleotide probe specificity. Where the used word size is not specified in the table, it is unknown.

'Contiguous Identity' refers to whether the tool uses a test for stretches of contiguous sequence identities with a non-target sequence in heuristics. Heuristics are often used for speed instead of thermodynamic calculations: Kane *et al.* find that for 50-mer oligonucleotide probes, *e.g.*, any contiguous sequence longer than 15 nt shared with a non-target indicates a significant chance of cross-hybridization (Kane *et al.*, 2000).

'Forward and Reverse Strand Match' refers to whether the oligonucleotide selection involves (or the tool can perform) similarity searches against both the forward strand and the reverse-complement to ensure that there are no cross-hybridizations to anti-sense transcripts (Lehner *et al.*, 2002; Lavorgna *et al.*, 2004; Yelin *et al.*, 2003).

'Percent Identity' refers to a heuristic use of the percentage of sequence identity between an oligonucleotide probe to a non-target sequence. Heuristics are often used for speed instead of thermodynamic calculations: For example, Kane *et al.* find that for 50-mer oligonucleotide probes a percentage similarity of greater than 75% to a non-target sequence target indicates a significant chance of cross-hybridization (Kane *et al.*, 2000), while He *et al.* suggest a sequence identity cut-off value of 85% for both 50-mer and 70-mer probes (He *et al.*, 2005).

SeqMatch is a custom sequence similarity search tool developed for YODA. The algorithm uses a word-based look-up approach with a minimum word size of 4 nucleotides for DNA sequences (*i.e.*, parameter **w = 4**).

Suffix Arrays/Trees allow an efficient sequence similarity search algorithm that exploits a sorted list of all the suffixes of a sequence to identify exact string searches (Manber and Myers, 1993). It takes $O(N \log N)$ time to build a suffix array, where N is the length of the sequence. A suffix array string search then completes in time $O(p + \log N)$, where p is the length of the sequence word.

'Target / Probe Mismatch Pos.' refers to whether the oligonucleotide tool takes into account the impact of mismatch positions between the target and probes as discussed by Hughes *et al*. Mismatches located toward the solution end (rather than the tethered end) of the probe significantly reduce signal intensity (Hughes *et al.*, 2001).

Design Tool	**GC Content**	**Free Energy**	**T_m Method**	**T_m Range**	**Non-specific Hybridization**	**Secondary Structure**	**Di-mer**	**Hair-pin**
ArrayOligoSelector	Yes	?	NN (unknown)	No	Yes	SW	Yes	?
GoArrays	No	?	NN; ***SL98***	Yes	No	MFOLD	?	?
OligoArray	Yes	Yes	NN; ***SL98***	Yes	Yes	MFOLD	Yes	Yes
OliCheck	?	?	Yes (unknown)	?	No	?	?	?
Oligodb	Yes	No	NN; `melting`	?	No	MFOLD	Yes	Yes
OligoDesign	No	No	NN; ***SL98***	No	No	Nussinov	Yes	?
OligoPicker	No	No	GC; Schildkraut	Yes	No	BLAST	Yes	Yes
OligoWiz	No	Yes	NN (unknown)	Yes	No	Yes (unknown)	Yes	?
Oliz	Yes	No	Yes (unknown)	Yes	No	No	No	No
Osprey	No	Yes	NN; Borer; ***SL98***	Yes	Yes	MFOLD	Yes	Yes
Picky	Yes	?	NN; SL96	Yes	Yes	Yes (unknown)	Yes	Yes
PRIMEGENS	Primer3	?	Breslauer	No	?	Primer3	Yes	?
PROBESEL	No	No	NN; ***SL98***	No	No	No	No	No
ProbeSelect	Yes	Yes	NN; ***SL98***	No	No	No	No	?
Promide	No	No	NN; ***SL98***	Yes	No	Yes (unknown)	Yes	No
ROSO	Yes	Yes	NN; ***SL98***	Yes	No	Yes (unknown)	Yes	Yes
YODA	Yes	?	NN; ***SL98***	Yes	No	Yes (unknown)	Yes	?

Table 4: Features of Selected Freely Available Oligonucleotide Probe Design Tools (II). *Legend:* over.

Legend for Table 4 (in column order):

'GC Content' refers to the probe composition being used for heuristics. It has been suggested that oligonucleotide probes containing between 30%–70% of guanine (G) and cytosine (C) nucleotides might be preferable (Kane *et al.*, 2000).

Primer3 is an oligonucleotide primer design tool available as a stand-alone software package and online (Rozen and Skaletsky, 2000). Primer3 calculates oligonucleotide melting temperature according to Breslauer (Rychlik and Rhoads, 1989; Breslauer *et al.*, 1986).

'Free Energy' refers to whether the tool calculates the Gibbs free energy ΔG of the probe-target duplex. While this can also be used to calculate the melting temperature T_m (see ***T_m* Method**) the binding energy between probe and target can directly be used as a measure of duplex stability. Little cross-hybridization was, *e.g.*, observed for 50-mer probes and non-targets with minimal binding free energies of more than –30 kcal/mol (He *et al.*, 2005); similarly for 70-mer probes with minimal binding free energies of more than –40 kcal/mol. The Gibbs free energy is discussed in `http://www.2ndlaw.com/gibbs.html`. Further background reading is found in the chapter `http://scholar.chem.nyu.edu/0651/notes/pchem/node55.html`.

'T_m* Method'** refers to whether the oligonucleotide selection tool calculates the melting temperature T_m of the probe–oligonucleotide duplex. Melting temperature is often used to characterize and compare the thermodynamic behaviour of probe candidates. As a full calculation is difficult, two-state approximations and semi-empirical approaches are typically employed (DeVoe and Tinoco, 1962; Gray and Tinoco, 1970; Uhlenbeck *et al.*, 1973; Tinoco *et al.*, 1973; Borer *et al.*, 1974; SantaLucia *et al.*, 1996; Allawi and SantaLucia, 1997; SantaLucia, 1998). It needs to be emphasized that extensive reviews have shown consistently superior performance of methods employing modern parameters (SantaLucia, 1998). Results from older approaches must therefore be deemed unreliable. Probe design tools that use the up-to-date unified parameters of SantaLucia (1998) are marked ***SL98 in the table. See section on melting temperature in this supplement.

***'T_m* Range'** refers to whether T_m of probe-candidates is thresholded. Allowing a wider range of T_ms provides a larger search space and hence gives more flexibility for finding specific probes, which is especially relevant for 'difficult' cases. On the other hand, many tools aim to provide a set of probes with uniform T_m. A user configurable range is meant to allow a trade-off between these aims.

'Non-specific Hybridization' refers to whether the cross-hybridization potential of an oligonucleotide candidate with all its non-targets are calculated. This is necessary for the selection of specific probes.

'Secondary Structure' refers to whether the tool tries to predict potential stable secondary structures that the oligonucleotide probe may form (self-hybridization / folding). As algorithms are based on sequence alignments, several tools employ standard methods like **BLAST** (Altschul *et al.*, 1997) or the Smith-Waterman (**SW**) sequence alignment algorithm (Smith and Waterman, 1981). Then there are more specialized approaches like the **Nussinov** algorithm (Nussinov *et al.*, 1990) and more complex algorithms that are typically implemented through external tools such as **Primer3** (Rozen and Skaletsky, 2000), **MFOLD** (Zuker, 2003) or **HyTher** (`http://ozone3.chem.wayne.edu/loginPage.html`). The latter two employ advanced models and calculations for secondary structure prediction taking into account a variety of folding possibilities.

'Dimer' refers to whether the tool makes any calculations to predict dimerization of the oligonucleotide probe. This is a special case of a secondary probe structure.

Hairpin' refers to whether the tool makes any calculations to predict hairpins within the oligonucleotide probe. This is a special case of a secondary probe structure.

Design Tool	Oligo Binding Pos.	Optimized Probe Len	Prohibited Motifs	Probes / Target	Target Regions / Probe	Exon / Intron Structure
ArrayOligoSelector	Yes (3')	No	Yes	?	1	No
GoArrays	?	No	Yes	?	2x 25-mer	No
OligoArray	Yes (3')	Yes	Yes	> 1	1	No
OliCheck	?	?	?	?	1	No
Oligodb	Yes	No?	Yes	1	1	No
OligoDesign	?	?	?	?	1	No
OligoPicker	Yes (Protein coding seq. only)	No	Yes	> 1	1	No
OligoWiz	Yes	Yes	Yes	> 1	1	Yes
Oliz	3' UTR option	No	No	1 ?	1	No
Osprey	Yes (5')	Yes	Yes	?	1	No
Picky	?	Yes	?	?	1	No
PRIMEGENS	?	?	?	?	1	No
PROBESEL	No	Yes	No	?	1	No
ProbeSelect	No	No	Yes	?	1	No
Promide	No	No	No	1	1	No
ROSO	Yes	No	Yes	> 1	1	No
YODA	Yes	No	Yes	?	1	No

Table 5: Features of Selected Freely Available Oligonucleotide Probe Design Tools (III). *Legend:* over.

Legend for Table 5 (in column order):

'Oligo Binding Pos.' refers to whether the tool allows influencing probe selection by the location of the probe target region along the target sequence. This is an important because, depending on the target labelling protocols employed, probe position will affect signal intensity. For example, the reverse transcriptase enzyme copies from the 3'-end of the target sequence, not necessarily progressing along the full-length of the target sequence and thus producing a 3' bias. In general, therefore, if poly-dT priming is used then probes should be close to the 3' end of the sequence. Conversely, if random priming is used then probes should be towards the 5' end or the centre of the sequence because the likelyhood of a random primer initiating a copy that includes the 3'-terminal is relatively small.

'Optimized Probe Len' refers to whether the tool will adapt oligonucleotide length. Non-uniformity in oligonucleotide length can, for example, achieve greater uniformity thermodynamic properties and reduce the chance of cross-hybridization.

'Prohibited Motifs' refers to whether the tool allows the user to provide specific sequence motifs that must be avoided in selected probes. This can, for example, be used to avoid low-complexity regions such as long stretches of mono- and di-nucleotide bases if the tool does not detect these already.

'Probes / Target' refers to whether multiple probes can be designed per target.

'Target Regions / Probe' refers to whether composite probes that bind multiple regions of a target are supported. To our knowledge, there is at present only one tool taking such an approach, GoArrays [38], which-constructs probe oligonucleotides with two contiguous 25-mer sequences specific to the target linked by random sequences of up to 6 nt to create an oligonucleotide probe of around 55 nt in length. A stable hybridization between the composite probe and the cDNA target can be achieved by the formation of a loop. Such a 'split' of longer probes can provide extra flexibility in avoiding non-specific hybridization to related targets without loosing the higher binding energy and thus sensitivity of longer probes.

'Exon / Intron Structure' refers to whether a tool allows the design of probes to distinguish between splice variants specifically validating exon / intron structure.

References

Allawi, H. T., and SantaLucia, J., Jr. (1997). Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* **36,** 10581-94.

Allawi, H. T., and SantaLucia, J., Jr. (1998a). Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry* **37,** 2170-9.

Allawi, H. T., and SantaLucia, J., Jr. (1998b). Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* **37,** 9435-44.

Allawi, H. T., and SantaLucia, J., Jr. (1998c). NMR solution structure of a DNA dodecamer containing single G*T mismatches. *Nucleic Acids Res* **26,** 4925-34.

Allawi, H. T., and SantaLucia, J., Jr. (1998d). Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Res* **26,** 2694-701.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215,** 403-10.

Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389-3402.

Borer, P. N., Dengler, B., Tinoco, I., Jr., and Uhlenbeck, O. C. (1974). Stability of ribonucleic acid double-stranded helices. *J Mol Biol* **86,** 843-53.

Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., and DeRisi, J. L. (2003). Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol* **4,** R9.

Breslauer, K. J., Frank, R., Blocker, H., and Marky, L. (1986). Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A* **83,** 3746-3750.

Charbonnier, Y., Gettler, B., Francois, P., Bento, M., Renzoni, A., Vaudaux, P., Schlegel, W., and Schrenzel, J. (2005). A generic approach for the design of whole-genome oligoarrays, validated for genomotyping, deletion mapping and gene expression analysis on Staphylococcus aureus. *BMC Genomics* **6,** 95.

Chen, H., and Sharp, B. M. (2002). Oliz, a suite of Perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3' untranslated region. *BMC Bioinformatics* **3,** 27.

Chou, H. H., Hsia, A. P., Mooney, D. L., and Schnable, P. S. (2004). Picky: oligo microarray design for large genomes. *Bioinformatics* **20,** 2893-902.

DeVoe, H., and Tinoco, I., Jr. (1962). *J Mol Biol* **4,** 500-517.

Gordon, P. M., and Sensen, C. W. (2004). Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays. *Nucleic Acids Res* **32,** e133.

Gray, D. M., and Tinoco, I., Jr. (1970). *Biopolymers* **9,** 223-244.

He, Z., Wu, L., Li, X., Fields, M. W., and Zhou, J. (2005). Empirical establishment of oligonucleotide probe design criteria. *Appl Environ Microbiol* **71,** 3753-60.

Howley, P. M., Israel, M. A., Law, M. F., and Martin, M. A. (1979). A rapid method for detecting and mapping homology between heterologous DNAs. Evaluation of polyomavirus genomes. *J Biol Chem* **254,** 4876-83.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P.,

Friend, S. H., and Linsley, P. S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19,** 342-7.
Kaderali, L., and Schliep, A. (2002). Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics* **18,** 1340-9.
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.
Koehler, R. T., and Peyret, N. (2005). Effects of DNA secondary structure on oligonucleotide probe binding efficiency. *Computational Biology and Chemistry* **29,** 393-397.
Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C. M., and Casari, G. (2004). In search of antisense. *Trends Biochem Sci* **29,** 88-94.
Le Novere, N. (2001). MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics* **17,** 1226-7.
Lehner, B., Williams, G., Campbell, R. D., and Sanderson, C. M. (2002). Antisense transcripts in the human genome. *Trends Genet* **18,** 63-5.
Li, F., and Stormo, G. D. (2001). Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* **17,** 1067-76.
Manber, U., and Myers, E. W. (1993). Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.* **22,** 935-948.
Mrowka, R., Schuchhardt, J., and Gille, C. (2002). Oligodb--interactive design of oligo DNA for transcription profiling of human genes. *Bioinformatics* **18,** 1686-7.
Nielsen, H. B., Wernersson, R., and Knudsen, S. (2003). Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res* **31,** 3491-6.
Nordberg, E. K. (2005). YODA: selecting signature oligonucleotides. *Bioinformatics* **21,** 1365-70.
Nussinov, R., Shapiro, B., Le, S., and Maizel, J. V. (1990). Speeding up the dynamic algorithm for planar RNA folding. *Math. Biosci.* **100,** 33-47.
Peyret, N., Seneviratne, P. A., Allawi, H. T., and SantaLucia, J., Jr. (1999). Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry* **38,** 3468-77.
Rahmann, S. (2003). Fast large scale oligonucleotide selection using the longest common factor approach. *J Bioinform Comput Biol* **1,** 343-61.
Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G., and Fayard, J. M. (2004). ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics* **20,** 271-3.
Rimour, S., Hill, D., Militon, C., and Peyret, P. (2005). GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics* **21,** 1094-103.
Rouillard, J. M., Herbert, C. J., and Zuker, M. (2002). OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18,** 486-7.
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.
Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132,** 365-86.

Rychlik, W., and Rhoads, R. (1989). A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucleic Acids Res* **17,** 8543-8551.

SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings- National Academy of Sciences USA* **95,** 1460-1465.

SantaLucia, J., Jr., Allawi, H. T., and Seneviratne, P. A. (1996). Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35,** 3555-62.

Schildkraut, C. (1965). Dependence of the melting temperature of DNA on salt concentration. *Biopolymers* **3,** 195-208.

Shai, O., Frey, B. J., Morris, Q. D., Pan, Q., Misquitta, C., and Blencowe, B. J. (2004). Probabilistic Inference of Alternative Splicing Events in Microarray Data. *In* "18th Annual Conference on Neural Information Processing Systems" (L. K. Saul, Y. Weiss, and L. Bottou, Eds.), Vol. 17. Neural Information Processing Systems Foundation.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195-197.

Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. (1996). Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* **24,** 4501-5.

Tinoco, I., Jr., Borer, P. N., Dengler, B., Levin, M. D., Uhlenbeck, O. C., Crothers, D. M., and Bralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol* **246,** 40-1.

Tolstrup, N., Nielsen, P. S., Kolberg, J. G., Frankel, A. M., Vissing, H., and Kauppinen, S. (2003). OligoDesign: Optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling. *Nucleic Acids Res* **31,** 3758-62.

Uhlenbeck, O. C., Borer, P. N., Dengler, B., and Tinoco, I., Jr. (1973). Stability of RNA hairpin loops: A 6 -C m -U 6. *J Mol Biol* **73,** 483-96.

Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T., and Itakura, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res* **6,** 3543-57.

Wang, X., and Seed, B. (2003). Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* **19,** 796-802.

Wernersson, R., and Nielsen, H. B. (2005). OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res* **33,** W611-5.

Xu, D., Xu, Y., Li, G., Zhou, J. (2000). A Computer Program for Generating Gene-Specific Fragments for Microarrays. *In* "Currents in Computational Molecular Biology" (R. S. a. T. T. S. Miyano, Ed.), pp. 3-4. Universal Academy Press, Inc., Tokyo, Japan.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K., and Rotman, G. (2003). Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21,** 379-86.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31,** 3406-15.