Thermodynamics of competitive surface adsorption on DNA microarrays

2006 J. Phys.: Condens. Matter 18 S491

(http://iopscience.iop.org/0953-8984/18/18/S02)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **18** (2006) S491–S523 doi:10.1088/0953-8984/18/18/S02

Thermodynamics of competitive surface adsorption on DNA microarrays

Hans Binder

Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Haertelstraße 16-18, Germany

E-mail: binder@izbi.uni-leipzig.de

Received 24 July 2005, in final form 11 December 2005
Published 19 April 2006
Online at stacks.iop.org/JPhysCM/18/S491

Abstract

Gene microarrays provide a powerful functional genomics technology which permits the expression profiling of tens of thousands of genes in parallel. The basic idea of their functioning is based on the sequence specificity of probe–target interactions combined with fluorescence detection. In reality, this straightforward principle is opposed by the complexity of the experimental system due to imperfections of chip fabrication and RNA preparation, due to the non-linearity of the probe response and especially due to competitive interactions which are inherently connected with the high throughput character of the method. We theoretically analysed aspects of the hybridization of DNA oligonucleotide probes with a complex multicomponent mixture of RNA fragments, such as the effect of different interactions between nucleotide strands competing with the formation of specific duplexes, electrostatic and entropic blocking, the fragmentation of the RNA, the incomplete synthesis of the probes and 'zipping' effects in the oligonucleotide duplexes. The effective hybridization affinities of microarray probes are considerably smaller than those for bulk hybridization owing to the effects discussed, but they correlate well with the bulk data on a relative scale. In general, the hybridization isotherms of microarray probes are shown to deviate from a *Langmuir*-type behaviour. Nevertheless isotherms of the *Langmuir* or *Sips* type are predicted to provide a relatively simple description of the non-linear, probe-specific concentration dependence of the signal intensity of microarray probes.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Gene microarrays provide a powerful functional genomics technology which permits the expression profiling of tens of thousands of genes in parallel [1, 2]. Current applications range from global analyses of transcriptional programmes of different organisms [3, 4] over

the establishment of novel criteria for the classification and prognostics of diseases [5–7] to the accelerated discovery of drug targets [8, 9]. Moreover, the actual technology enables, at least in principle, very detailed genome analysis with an interrogation resolution of a few subsequent base pairs using so-called tiling arrays [10].

The working principle of the microarray technique is based on formation of duplexes (hybridization) between target RNA extracted from cell lines or tissues on one hand and complementary DNA nucleotide strands grafted to the chip (the probe molecules) on the other hand. The target includes mRNA, which transcribes genetic information into proteins, but potentially also 'non-coding' RNA, which, for example, can regulate gene expression [11]. From several thousands up to a few millions of different DNA oligonucleotide sequences can be immobilized on a support such as glass, silicon or nylon membrane in a grid of probe spots. Each probe spot ideally consists of oligomers of one sequence if one neglects fabrication errors. It is therefore representative for a certain genomic sequence and probes the abundance of the respective, complementary RNA transcript.

Duplexes formed can be detected using different labelling techniques such as applying fluorescent and radionucleotide markers or quantum dots. The integral signal of each probe spot is related to the amount of bound RNA, which in turn serves as a measure of the degree of expression of the respective gene in terms of the concentration of complementary RNA in the sample solution used for hybridization.

Different kinds of DNA arrays are designed for RNA profiling, which differ in the type of the probe (cDNA or synthetic oligonucleotides) and in the DNA density on the array (see e.g. [12]). So called high density oligonucleotide arrays (HDONA) are produced by a photolithographic technology, which allows the synthesis of oligonucleotide sequences on the chip surface in an extremely high density. In this way about 10^5–10^6 different probe spots relating to up to 50 000 genes can be localized on one microarray of an area of about one square centimetre [13].

The factors that control hybridization have been extensively investigated for DNA/RNA duplex formation in bulk solution (see, e.g., [14–18]). In contrast, fewer investigations have focused on the *in situ* kinetics and thermodynamics for surface immobilized probes interacting with solution phase targets, where the molecular level processes are more complex (see, e.g., [19–28]). Hybridization on microarrays is apparently governed by an intricate interplay between effects such as the stability of RNA/DNA duplexes, surface adsorption to a heterogeneous ensemble of binding sites, surface electrostatics and diffusion, fluorescence emission and also non-equilibrium thermodynamics [22, 23, 25–27, 29–34]. The importance of such complex factors is increasingly recognized but remains unaddressed in many respects in published theoretical and experimental work.

Physical models open up the possibility of adapting well proven concepts of thermodynamics, statistical mechanics, molecular physics and chemical kinetics to the problem of surface hybridization on microarrays and of explaining the observed probe signal in terms of a set of well defined experimental parameters. Finally, this approach shows a great potential for relating the probe signal to the underlying RNA concentration and in this way to improve existing expression measures based on statistical models. Recent work mainly explores selected physical aspects of microarray hybridization such as the origin of non-linearities in the probe responses and sequence effects in the behaviour of perfectly matched and mismatched probes [25, 29, 30, 23, 22, 35–41]. Other studies developed models which describe the signal of the probes as a function their sequence [29, 42, 32] partly with direct relations to the free energy for RNA/DNA duplex formation [31]. Despite the progress achieved in the understanding of surface hybridization it seems that the system producing the measured intensities is too complex to be currently fully described by a relatively simple physical model.

The present paper theoretically analyses selected aspects of the hybridization of DNA oligonucleotide probes in a complex environment which is provided by a multicomponent mixture of RNA fragments in terms of simple models. In the first two sections we discuss different processes which compete with formation of specific duplexes between probe and target and estimate the relations between the respective equilibrium constants using a simple interaction model for the dimerization of oligonucleotide strands. Consequences of electrostatics and polymeric flexibility, of the fragmentation of the RNA, of the incomplete synthesis of the probes and of 'zipping' effects in the oligonucleotide duplexes for the hybridization efficiency of the microarray probes are analysed in the subsequent sections. The results of this theoretical study illustrate the basic trends arising from the effects considered in a rather qualitative fashion. This approach also aims at establishing the tools for the quantitative analysis of microarray data in terms of an appropriate hybridization isotherm which links the measured probe signal with the RNA concentration in the sample solution. In the accompanying paper we apply this isotherm to study experimental microarray data [43].

2. Competing interactions upon microarray hybridization

The microarray experiment aims at estimating the amounts of specific duplexes, P–S, of surface grafted probes and 'specific' complementary target RNA, S, as a measure of the total concentration of the latter species in the sample solution. The concentration of the specific dimers (as indicated by the brackets, i.e., [P–S]) is related to the concentration of free species according to the mass action law

$$[\mathrm{P-S}] = K^{\mathrm{P-S}} \cdot [\mathrm{S}^{\mathrm{free}}] \cdot [\mathrm{P}^{\mathrm{free}}], \tag{2.1}$$

where $K^{\mathrm{P-S}}$ is the bimolecular reaction constant of duplex formation. Only 'freely' accessible, i.e., monomeric and unfolded, target RNA ($\mathrm{S}^{\mathrm{free}}$) and DNA probe oligomers ($\mathrm{P}^{\mathrm{free}}$) can assemble into dimers. It is important to take into account that the formation of probe/target duplexes competes with other molecular interactions in the typical set-up of the microarray experiment (see figure 1 for illustration): (i) non-specific hybridization of the probe with RNA fragments which partly match the probe sequence via complementary WC pairs (P–NS); (ii) intramolecular folding of target RNA which makes the complementary region of the target partly inaccessible for duplex formation (S-fold); (iii) intramolecular folding of the probes (P-fold); (iv) bulk dimerization of the target in terms of heterodimers with RNA fragments which partly match the target sequence (S–NS); and (v) bulk dimerization of the target in terms of partly self-complementary homodimers (S–S).

These additional interactions effectively reduce the amount of free RNA and of probe oligonucleotides compared with the total amount of target RNA in the solution (S) and of the probe oligonucleotides on the chip (P), respectively. In other words, the probe and the target must first dissociate into unfolded monomers before they can form a P–S duplex. The analogous scheme applies to the non-specific transcripts with folded monomers (NS-fold) and dimerized homoduplexes and heteroduplexes, (NS–NS and NS–NS′, respectively).

2.1. Non-specific hybridization

The sample solution used for hybridization usually contains a very heterogeneous cocktail of different RNA fragments with a broad distribution of sequences and lengths. These non-specific RNA fragments can bind in significant amounts to the probes and in this way reduce the number of accessible binding sites for specific target RNA according to

$$\sum_k [\mathrm{P{-}NS}_k] = \sum_k K_k^{\mathrm{P-NS}} \cdot [\mathrm{NS}_k^{\mathrm{free}}] \cdot [\mathrm{P}^{\mathrm{free}}]. \tag{2.2}$$

In practice it seems impossible to determine the concentrations ([NS_k]) and binding constants (K_k^{P-NS}) of all relevant non-specific fragments in an appropriate fashion. Their 'total' hybridization strength and the amount of non-specific hybridization of the probes are

$$X^{P-NS} = K^{P-NS} \cdot [NS] = \sum_k K_k^{P-NS} \cdot [NS_k^{free}] \qquad \text{and}$$
$$[P{-}NS] = \sum_k [P{-}NS_k], \tag{2.3}$$

respectively. The decomposition of X^{P-NS} into an effective equilibrium constant of non-specific hybridization, K^{P-NS}, and the effective total concentration of non-specific transcripts, [NS], replaces the equilibria of binding of the probe with all relevant non-specific RNA sequences by one equilibrium of the probe with one average, 'characteristic' fragment of non-specific transcripts. In other words, the cocktail of non-specific RNA fragments is assumed to act like a single species in accordance with previous treatments of cross-hybridization [22]. The effect of an ensemble of non-specific RNA fragments of different lengths and sequences will be analysed in more detail below in terms of a simple microscopic model.

2.2. *Folding of the probes*

The unimolecular folding reaction effectively reduces the amounts of probes which are accessible for surface hybridization according to [39, 44]

$$[P^{free}] = \frac{[P] - [P-S]}{(1 + K^{P\text{-fold}})}. \tag{2.4}$$

2.3. *Dimerization and folding of the RNA fragments*

Bulk dimerization and folding of the target decreases the effective concentration of specific transcripts in solution. In analogy with equation (2.3) we replace the dimerization equilibria of the target with all relevant NS fragments by

$$K^{S-NS} \cdot [NS] = \sum_k K_k^{S-NS} \cdot [NS_k^{free}]. \tag{2.5}$$

The concentration of free target becomes in the limit of large excess of dissolved RNA ([S], $[S^{free}] \gg [P-S], [P]$)

$$[S^{free}] \approx \frac{[S]}{(1 + K^{S\text{-fold}} + K^{S-NS}[NS] + K^{S-S}[S^{free}])}. \tag{2.6}$$

In the special cases of weak and strong affinity for homodimerization (compared with the other terms in the denominator of equation (2.6)) one obtains

$$[S^{free}] \approx \frac{[S]}{(1 + K^{S\text{-fold}} + K^{S-NS}[NS])}; \qquad K^{S-S}[S] \ll 1$$
$$\text{and} \qquad [S^{free}] \approx \sqrt{[S]/K^{S-S}}; \qquad K^{S-S}[S] \gg 1, \tag{2.7}$$

respectively. Hence, the free concentration of specific transcripts becomes a non-linear function of the total target concentration [S] in the latter case (see also [22, 44]).

The reaction scheme shown in figure 1 analogously applies also to the non-specific RNA fragments with respect to folding and dimerization reactions. One can therefore rewrite equation (2.6) accordingly as

$$[NS^{free}] \approx \frac{[NS]}{(1 + K^{NS\text{-fold}} + K^{S-NS}[S] + K^{NS-NS'|}[NS])}, \tag{2.8}$$

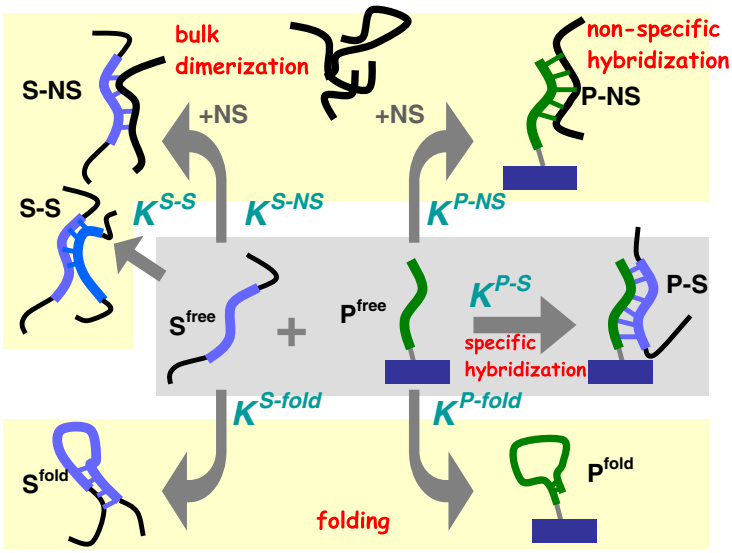

Figure 1. The scheme illustrates the competing interactions on microarrays. The specific hybridization between the RNA target and the DNA probe is shown in the central part. Its yield is decreased by bulk dimerization, non-specific hybridization and intramolecular folding of the probe and target.

if one ignores homodimerization. The last term in the dominator considers heterodimers between different non-specific fragments. Equation (2.8) further assumes a sufficient large excess of non-specific RNA fragments, [NS] $\gg$ [S], which prevents the concentration dependent depletion of [NS].

2.4. The binding isotherm and overall binding constants

Insertion of equations (2.4), (2.6) and (2.8) into equation (2.1) and rearrangement provides the 'surface coverage', θ, as a function of the concentrations of the species considered and the respective equilibrium constants (see figure 1 for designations) in the limit of large excess of dissolved RNA [39],

$$\begin{aligned} \theta \equiv \frac{[\mathrm{P}-\mathrm{S}]}{[\mathrm{P}]} \approx \frac{X}{1+X} \qquad & \text{with } X = X^{\mathrm{S}} + X^{\mathrm{NS}}, \\ X^{\mathrm{S}} = K^{\mathrm{S}} \cdot [\mathrm{S}] \qquad \text{and} \qquad & X^{\mathrm{NS}} = K^{\mathrm{NS}} \cdot [\mathrm{NS}] \end{aligned} \tag{2.9}$$

with

$$\begin{aligned} K^{\mathrm{S}} &\approx \frac{K^{\mathrm{P-S}} \cdot (1 + K^{\mathrm{P\text{-}fold}})^{-1}}{(1 + K^{\mathrm{S\text{-}fold}} + K^{\mathrm{S-NS}}[\mathrm{NS}] + \sqrt{K^{\mathrm{S-S}}[\mathrm{S}]})} \qquad \text{and} \\ K^{\mathrm{NS}} &\approx \frac{K^{\mathrm{P-NS}} \cdot (1 + K^{\mathrm{P\text{-}fold}})^{-1}}{(1 + K^{\mathrm{NS\text{-}fold}} + K^{\mathrm{S-NS}}[\mathrm{S}] + K^{\mathrm{NS-NS'}}[\mathrm{NS}])}. \end{aligned} \tag{2.10}$$

Equation (2.9) represents a two-species *Langmuir*-type adsorption isotherm. It links the 'surface coverage' of the adsorbent, θ (defined as the fraction of occupied DNA probes on the chip) with the effective 'adsorption strength', X^{h} (h = S, NS), of the sorbates in the supernatant solution (i.e., S- and NS-RNA). Equation (2.9) assumes that the surface reaction virtually does not deplete the total amount of RNA in the 'reservoir' provided by the hybridization solution ([S], [NS] $\gg$ [P − S], [P]). The equation for K^{S} in equation (2.10) was chosen to meet the

limiting cases at small and high [S] in a simple way (see equation (2.7)). It progressively decreases upon increasing target concentration owing to target–target dimerization.

Equations (2.9) and (2.10) show that surface hybridization on microarrays is governed by two effective, 'apparent' hybridization constants, $K_{\mathrm{app}}^{\mathrm{P-S}} = K^{\mathrm{S}}$ and $K_{\mathrm{app}}^{\mathrm{P-NS}} = K^{\mathrm{NS}}$, which are (i) directly related to the respective intrinsic hybridization constants, $K^{\mathrm{P-S}}$ and $K^{\mathrm{P-NS}}$; and (ii) inversely related to the equilibrium constants of several competing reactions in the system which effectively decrease the hybridization yield of the probe–target dimerization. The apparent hybridization constants are consequently reduced by a factor $K_{\mathrm{comp}}^{\mathrm{h}} < 1$ according to

$$K^{\mathrm{h}} = K^{\mathrm{P-h}} \cdot K_{\mathrm{comp}}, \tag{2.11}$$

with $\mathrm{h} = \mathrm{S}, \mathrm{NS}$ and $K_{\mathrm{comp}} \approx K_{\mathrm{comp}}^{\mathrm{S}} \approx K_{\mathrm{comp}}^{\mathrm{NS}}$. In the absence of competing processes ($K^{\text{S-fold}} \approx K^{\text{P-fold}} \approx K^{\mathrm{S-NS}}[\mathrm{NS}] \approx K^{\mathrm{S-NS}}[\mathrm{S}] \approx K^{\mathrm{S-S}}[\mathrm{S}] \approx K^{\mathrm{NS-NS}}[\mathrm{NS}] \ll 1$) the constants in the isotherm are given by the intrinsic affinities for P–S and P–NS duplex formation, i.e., $K^{\mathrm{S}} \approx K^{\mathrm{P-S}}$ and $K^{\mathrm{NS}} \approx K^{\mathrm{P-NS}}$, respectively.

2.5. Interactions on the chip surface

The scheme shown in figure 1 neglects direct interactions between surface fixed species, namely the 'bridging' between neighbouring free probe oligomers (P–P) and between the dangling ends of RNA fragments bound to adjacent probes (PS–SP, PNS–NSP, PS–NSP) and the respective cross-terms (PNS–P, PS–P). In particular, formation of duplexes between free probe oligomers decreases the number of available free binding sites for P–S interactions and thus the effective affinity constant K^{S} (see equation (2.10)). The interactions between the dangling ends of bound NS-RNA are expected to increase the stability of the bound state relatively to the free one with similar consequences for K^{S}.

The interactions between the dangling tails of bound RNA are related to the reaction constants for RNA–RNA dimerization and folding in bulk solution (S–S, S–NS, NS–NS, NS–NS′, S-fold, NS-fold). Consequently their effect can be partly included in the bulk terms considered in scheme 1. The same argument holds for the probe–probe self-interaction propensity which is related to the affinity for folding (P-fold). It turns out, however, that especially the latter interactions have only a tiny effect on the overall reaction rate due to the relatively small number of base pairings available in the folded structures (see below). It has recently been suggested that, compared with a complementary target, the presence of a mismatch might facilitate bridging by destabilizing duplex formation at the location of the bridge [45].

Note that the interactions on the chip surface increase with decreasing distance between neighbouring probe oligomers. They are consequently a function of the two-dimensional density of the oligomers on the microarray which is not considered in equation (2.10). In addition to the 'chemical' base pair interactions which stabilize the complexes discussed, repulsive coulombic and entropic forces hamper the binding of transcripts to the probes. The charge density of the surface and the crowding of surface grafted oligomers amplify with increasing number of probe strands per unit area and thus they are also functions of the surface density of the probes. The mechanism and the consequences of surface related electrostatic and entropic blocking are discussed in a separate section below.

3. Microscopic model of interacting oligonucleotide strands

In the next step we will evaluate the relevance of the different competitive reactions and estimate the relations between the respective reaction constants using a simple microscopic model of microarray hybridization. Microarrays are normally used to measure the degree of

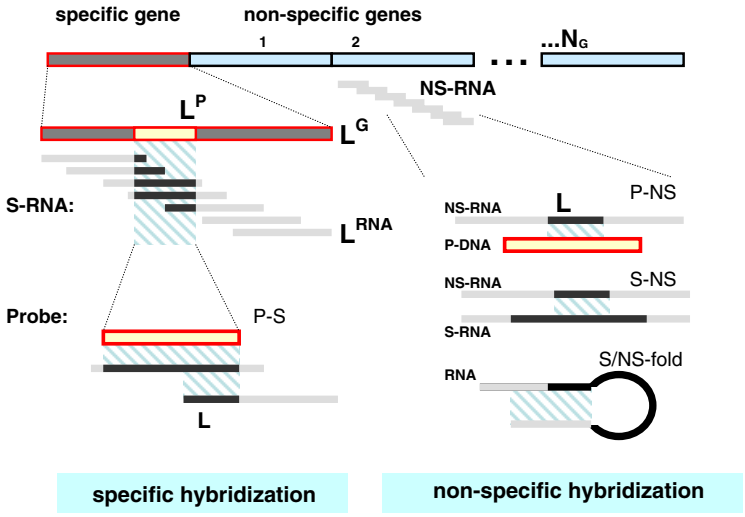

Figure 2. Interactions of a microarray probe with specific (S-) and non-specific (NS-) RNA fragments which are transcribed from the model genome composed of one 'specific' gene and N_G 'non-specific' genes. The probe sequence (ξ^P) of length L^P is a subsequence of the S gene. It binds the specific target RNA via complementary bases either over the full length of the probe or only partly over a sequence of L bases due to the fragmentation of the RNA into pieces of length L^{RNA} as indicated in the left part of the figure. The hatched area relates to the region of WC pairings. The right part of the scheme shows different complexes formed by the NS-RNA transcribed from the NS genes such as non-specific dimers of the probe (P–NS), of the target (S–NS) and folded species (see the text).

'expression' in terms of the mRNA transcribed from the genes to encode the respective proteins. The preparation technique for GeneChip experiments includes the extraction of total RNA from (many) cells of, ideally, one type, its conversion to cDNA and amplification. The amplified cDNA is reversely transcribed into cRNA, which is then fragmented and hybridized onto the chip. The hybridized chip is washed, stained with a fluorescing conjugate and scanned.

The *in vitro* transcription produces cRNA with a broad length distribution ranging from a few hundreds up to 4000–6000 nucleotides and with an average length of about 1000 nucleotides [46]. In the fragmentation step the cRNA is randomly cut into pieces with an average length about $\langle L^{RNA} \rangle \approx 100$ and a FWHM (full width at half-maximum) of the respective frequency distribution of about 100.

3.1. Model genome

The extracted RNA ideally reflects the 'transcriptome' expressed from the genome of a particular cell. It usually represents a complex, very heterogeneous mixture of RNA fragments of different sequences and lengths. Let us assume the following simple model to assess the relations between the different processes shown in figure 1. The transcribed genome consists of $N^G + 1$ different genes of uniform length of L^G 'coding' bases each (see figure 2). The bases are randomly distributed along the genes. A number of L^P subsequent bases taken from one of the genes forms the probe sequence, ξ^P, for binding complementary, 'specific' mRNA transcribed from the gene of interest and in this way to probe its abundance. All N^G other genes are consequently per definition non-specific with respect to the probe considered.

3.2. Interaction model of DNA and RNA complexes

The probe sequence, ξ^{P}, with P = PM usually perfectly matches the target sequence, ξ^{S}, of the specific RNA in terms of Watson–Crick (WC) pairings. Let us assume that the cRNA obtained for hybridization from the model genome is randomly fragmented into pieces of uniform length, L^{RNA}. These fragments can either match the probe over its full length L^{P}, or only partly over $L < L^{\mathrm{P}}$ complementary bases depending on the position of a particular cut which fragments the RNA at a position outside or inside of the target region, respectively (see figure 2). In this section we consider only specific RNA fragments with an intact target region. The more general case is discussed below.

The effective equilibrium constants, K^{C}, of the dimers considered (C = P–S, P–NS, S–NS, S–S, NS–NS′, NS–NS) and folded monomers (C = P-fold, S fold, NS fold) can be written in a general fashion as

$$K^{\mathrm{C}} \approx \sum_{L=1}^{L\,\max} P^{\mathrm{C}}_{\mathrm{match}}(L, L^{\mathrm{C}}_{\mathrm{res}}) \cdot N^{\mathrm{C}}_{\mathrm{match}}(L, L^{\mathrm{C}}_{\mathrm{max}}) \cdot K^{\mathrm{C}}_0(L) \qquad \text{with } K^{\mathrm{C}}_0(L) \equiv W^{\mathrm{C}}_0 \cdot \exp(-G^{\mathrm{C}}_0(L)/RT). \tag{3.1}$$

The sum in equation (3.1) runs over the microstates of the corresponding bimolecular or unimolecular complex C. These microstates can differ in the number, L, of WC pairings which stabilize the complex. The first 'probability' factor in equation (3.1), $P^{\mathrm{C}}_{\mathrm{match}}(L, L_{\mathrm{res}})$, considers the probability of finding a particular subsequence of length L within a 'sequence reservoir' of length L_{res}. This subsequence provides the complementary bases for complex formation. The second 'frequency' factor, $N^{\mathrm{C}}_{\mathrm{match}}(L, L_{\mathrm{max}})$, specifies the number of ways of arranging L pairings within a sequence region of length L_{max}, which at least partially includes the probe sequence, ξ^{P}, or its complementary, specific target sequence, ξ^{S}. Note that the product $P^{\mathrm{C}}_{\mathrm{match}}(L, L_{\mathrm{res}}) \cdot N^{\mathrm{C}}_{\mathrm{match}}(L, L_{\mathrm{max}})$ is the mean number of realizations for L pairings according to the binomial distribution. It weights the association constant for the corresponding complex, $K^{\mathrm{C}}_0(L)$, which is stabilized via L complementary WC base pairings. The binding constant is, in turn, related to the free energy of complex formation, $G^{\mathrm{C}}_0(L)$ (RT is the thermal energy). The cratic factor, W^{C}_0 (given in units of K^{C}_0), accounts for the change of ideal mixing entropy in the corresponding reaction (see [47], pp 283).

3.3. Stability of hybrid duplexes

The free energy of complex formation in general depends on the particular sequence of paired bases (see below). For simplicity we will assume a linear function of L, the number of WC pairings,

$$G^{\mathrm{C}}_0(L)/RT = g_{\mathrm{init}} + L \cdot g(x_{\mathrm{GC}}) \cdot f^{\mathrm{A/A'}} \qquad \text{with} \qquad g(x_{\mathrm{GC}}) = g_0 + x_{\mathrm{GC}} \cdot \Delta g_{\mathrm{GC}} + (1 - x_{\mathrm{GC}}) \cdot \Delta g_{\mathrm{AT}}. \tag{3.2}$$

The mean contribution per single pair, g_0, is modulated by an incremental term which considers, e.g., stronger interactions for G–c* and especially C–g pairings compared with T–a and A–u* in the hybrid duplexes (see below; upper case letters relate to the DNA probe, lower case letters to the RNA, the asterisk indicates labelling). $g_{\mathrm{init}} \approx -2g_0$ is an initiation term [15]. We further assume that the fraction of G and C, the so-called GC content x_{GC}, is uniformly distributed along the whole probe length L^{P}. The energetic contribution of non-WC pairings is ignored in this simple approach.

In the calculations we used the values $g_0 = 0.4$, $\Delta g_{\mathrm{GC}} = 0.2$ and $\Delta g_{\mathrm{AT}} = -0.2$. They relate to the free energy data for DNA/RNA duplex formation in solution taken from [15, 18]

which are however scaled down by the factor ∼0.3 or, equivalently, by an apparent temperature of ∼900 K, and averaged over the four nearest neighbours to obtain single-base related data. The reduction factor accounts for effects such as competitive complex formation and electrostatic and entropic blocking which are expected to decrease the strength of DNA/RNA interactions near surfaces compared with that in solution (see below).

Which is more stable among RNA/RNA (S–NS, S–S, NS–NS′, NS–NS, S-fold, NS-fold), DNA/RNA (P–S, P–NS) and DNA/DNA (P-fold) WC base pairings strongly depends on the sequence context [16]. On average, the interaction strength decreases with RNA/RNA > DNA/RNA > DNA/DNA [16, 48]. We therefore used an 'amplification' factor $f^{A/A'} = 1.2, 1.0, 0.8$ for $A/A' =$ RNA/RNA, DNA/RNA and DNA/DNA, respectively.

3.4. Specific duplexes

In this section we assume that the specific target binds over the whole probe length via complementary WC pairings. Hence, the specific P–S duplexes are characterized uniquely by one microstate with the association constant (insert into equation (3.1): $P^{P-S}_{match} = N^{P-S}_{match} = 1$ for $L = L^P$ and $P^{P-S}_{match} = 0$ otherwise)

$$K^{P-S}(L^P) = K^{P-S}_0(L^P) \equiv W^{P-S} \cdot \exp(-G^{P-S}_0(L^P)/RT). \tag{3.3}$$

Deviations from this simple relation owing to zipping effects, RNA fragmentation and truncated probes are discussed below.

3.5. Competing complexes

The NS fragments originate from N^G different non-specific genes which provide a reservoir of $L^{P-NS}_{res} = L^{tot} = N^G \cdot L^G$ nucleotide bases for non-specific P–NS duplexes. The probability that L of them match the probe sequence partly ($L < L^P$) or completely ($L = L^P$) by chance is

$$P^C_{match}(L, L^C_{res}) = 1 - \left(\frac{4^L - 1}{4^L}\right)^{L_{res}}. \tag{3.4}$$

Equation (3.4) provides a sigmoidal function which steeply decreases from $P_{match} \approx 1$ for $L < L^*$ to $P_{match} \approx 0$ for $L > L^*$. The critical length L^* depends only weakly on L^{tot}, the total length of the model genome. We arbitrarily set the total genome length to $L^{tot} = 10^7$ which relates, for example, to a number of translated genes and a length of their coding sequences of $N^G \sim 10^4$ and $L^G \sim 10^3$, respectively. In consequence, the maximum number of WC pairs in the non-specific hybrid duplexes effectively does not exceed $L^* \approx 13$ in our model calculation with $L^{tot} = 10^7$. Note that the variation of L^{tot} over several orders of magnitude between, e.g., 10^5 and 10^9, only weakly affects the critical length which increases from $L^* \sim 8.5$ to 15. Hence, the exact choice of L^{tot} is not crucial for the results obtained.

Also the bulk heteroduplexes (C = S–NS, NS–NS′) are recruited from the whole genome, i.e. $L^C_{res} = L^{tot}$. The sequence reservoir of the homoduplexes (C = S–S, NS–NS) is given by the length of the RNA fragments $L^C_{res} = L^{RNA} - 1$, which provides a relatively small critical number of matched bases of $L^* \approx 3.5$. The reservoir further reduces to $L^{RNA\text{-}fold}_{res} = L^{RNA} - L - 4$ for the folded RNA species (RNA-fold = S-fold, NS-fold) and to $L^{P\text{-}fold}_{res} = L^P - L - 4$ for the folded probes giving rise to $L^* < 3.5$ and 2.5, respectively. Here we assume that intramolecular folding consumes a loop of at minimum four bases which are excluded from WC pairings.

The number of ways to place L matches within a relevant sequence length of L_{max} is

$$N^C_{match}(L, L^C_{max}) = L^C_{max} - L + 1. \tag{3.5}$$

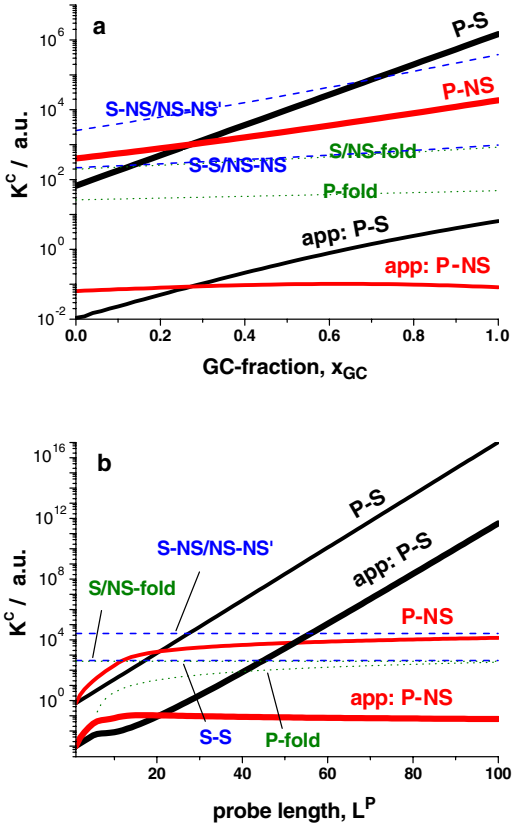

Figure 3. Reaction constants of the competitive processes as a function of the GC fraction (at $L^{\mathrm{P}} = 25$) and of the length of the probes (at $x_{\mathrm{GC}} = 0.5$). The curves are calculated using the interaction model equations (3.1)–(3.5) with the free energy parameters $g_0 = 0.4$, $\Delta g_{\mathrm{GC}} = 0.2$ and $\Delta g_{\mathrm{AT}} = -0.2$ in equation (3.2). The apparent hybridization constants of specific and non-specific hybridization are obtained by combination of the intrinsic constants using equation (2.10). Note the very similar behaviour of the intrinsic and of the apparent constants of P–S and P–NS complex formation despite the fact that the apparent values are considerably smaller.

The hybrid duplexes (C = P–h with h = S, NS) enable base pairings over the probe length, i.e., $L_{\mathrm{max}}^{\mathrm{P-h}} = L^{\mathrm{P}}$ whereas the RNA dimers (h–h′ and h–h) and the folded species are stabilized with at maximum $L_{\mathrm{max}}^{\mathrm{h-h'}} = L_{\mathrm{max}}^{\mathrm{h-h}} \approx L^{\mathrm{RNA}}$, $L_{\mathrm{max}}^{\mathrm{RNA\text{-}fold}} = (L^{\mathrm{RNA}} - L - 4)$ and $L_{\mathrm{max}}^{\mathrm{P\text{-}fold}} = (L^{\mathrm{P}} - L - 4)$ WC pairings, respectively.

3.6. Comparison of the association constants

Figure 3 compares the different association constants which are calculated by means of equation (3.1) as a function of the GC content (part (a)) and of the probe length (part (b)). The DNA/RNA related constant of specific binding, $K^{\mathrm{P-S}}$, increases linearly on a logarithmic scale as a function of x_{GC} and of L^{P} according to equation (3.2). The RNA/RNA related association constants of bulk dimerization and RNA folding are independent of the probe length. Their values indicate a considerably smaller affinity for homodimers (S–S, NS–NS) and folded fragments (S/NS-fold) compared with that for the heterodimers (S–NS, NS–NS′) owing to the larger sequence reservoir for complementary bases of the latter species (see above).

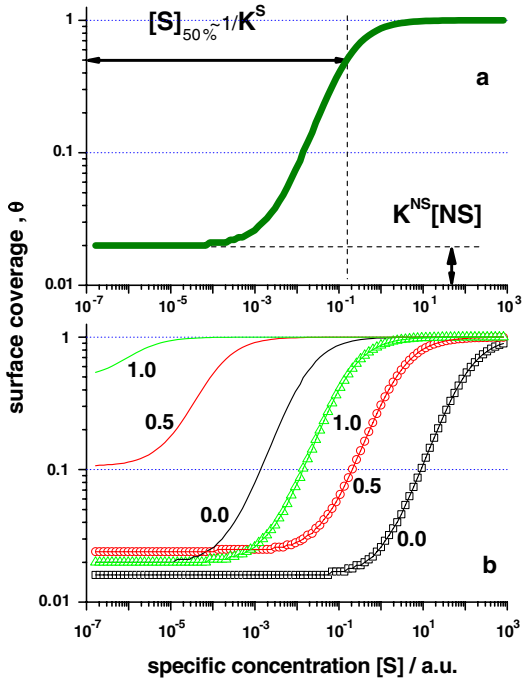

Figure 4. Hybridization isotherms of 25-meric probes which are calculated using equation (2.9) as a function of the specific transcript concentration (in arbitrary units). The inflection point at $[S]_{50\%}$ is inversely related to the binding constant for specific transcripts and defines the sensitivity of the probe (part (a)). The background level of non-specific hybridization, $K^{NS}[NS]$, affects its specificity (see the text). Part (b) shows the isotherms at varying GC content, x_{GC} (see values at the curves), which are calculated using either the intrinsic reaction constants, K^{P-S} and K^{P-NS} (thin lines), or their apparent values, K_{app}^{P-S} and K_{app}^{P-NS} (symbols; see also equation (2.10)). The latter curves are systematically shifted towards larger concentrations. Their non-specific background level is virtually independent of the GC content because non-specific binding is effectively compensated by other competing reactions (see also figure 3 and the text). The concentration of NS transcripts is $[NS] = 10^{-4}$ (thin curves) and $[NS] = 10^{-0}$ (symbols).

The constant of non-specific hybridization, K^{P-NS}, increases in a similar fashion or even more steeply with the probe length at $L^P < 20$, i.e. in the limit of short probes, compared with the association constant for specific binding, K^{P-S}. For longer probes, K^{P-NS}, however, asymptotically levels off into a constant value whereas K^{P-S} proceeds to increase. The behaviour of the former constant can be rationalized by the vanishing probability for $L > L^*$ WC pairing. In other words, the effective number of available WC pairings is limited by an upper value of $\sim L^*$ which prevents the further increase of K^{P-NS} at $L^P > L^*$. The affinity for probe folding, $K^{P\text{-fold}}$, increases in a similar asymptotic fashion as K^{P-NS} but its value is considerably smaller compared with the other binding constants due to the relatively small number of potentially available base pairings for backfoldings along the relatively short probe sequence.

It is important to note that the apparent association constants, K_{app}^{P-S} and K_{app}^{P-NS} (see equation (2.8)), behave in a similar fashion to the intrinsic hybridization constants, K^{P-S} and K^{P-NS}, upon varying L^P and x_{GC}. The apparent values are however considerably smaller by several orders of magnitude owing to the competitive processes.

3.7. Hybridization isotherms

Part (a) of figure 4 shows the binding isotherms which are calculated by means of equation (2.9) using the apparent association constants $K^{\mathrm{S}} = K_{\mathrm{app}}^{\mathrm{P-S}}$ and $K^{\mathrm{NS}} = K_{\mathrm{app}}^{\mathrm{P-NS}}$ (symbols) and the intrinsic bimolecular constants, $K^{\mathrm{S}} = K^{\mathrm{P-S}}$ and $K^{\mathrm{NS}} = K^{\mathrm{P-NS}}$ (see lines), upon varying the GC content of a 25-meric probe. The concentration value of the inflection point of the isotherms at half-coverage ($\theta = 0.5$) is inversely related to the affinity constant of specific hybridization, $[\mathrm{S}]_{50\%} = 1/K^{\mathrm{S}}$.

The isotherms relating to the apparent association constants are considerably shifted towards higher concentration values compared with the isotherms relating to the intrinsic hybridization constants. This shift reflects the reduction of the apparent values owing to competitive reactions (see figure 3). It was previously estimated that microarray probes indeed saturate only at a much higher concentration of the target RNA than one would expect from a simple target/probe equilibrium without consideration of the competitive processes [36, 49].

In the limit of small $[\mathrm{S}] \ll [\mathrm{S}]_{50\%}$ the isotherms level off into a plateau. It characterizes the limiting coverage due to non-specific hybridization, $\theta|_{[\mathrm{S}]\to 0} \approx X^{\mathrm{NS}} \propto K^{\mathrm{NS}}$ for $[\mathrm{NS}] =$ constant. The comparison of the two sets of curves reveals that the competitive interactions reduce the level of non-specific background. Most interestingly, this trend is paralleled by a considerable decrease of the variability of the limiting coverage upon varying the GC content. In other words, the background level becomes relatively insensitive for a particular probe sequence and for $[\mathrm{NS}] =$ constant in the presence of competitive interactions.

This result reflects the fact that the magnitude of $K^{\mathrm{P-NS}}$ is roughly comparable with that of the constants of the competing interactions and thus the concurring processes effectively compensate for the sequence specificity of $K^{\mathrm{P-NS}}$. In contrast, the hybridization constant for specific transcripts, $K^{\mathrm{P-S}}$, exceeds the reaction constants of the other processes by several orders of magnitude at $L^{\mathrm{P}} > L^*$ (see figure 3; $K^{\mathrm{P-NS}}$ asymptotically levels off at $L > L^*$) and thus it is sequence specificity is affected much more weakly by the competitive interactions. This different behaviour becomes evident also in part (a) of figure 3. The slope of $K_{\mathrm{app}}^{\mathrm{P-NS}}$ is distinctly reduced compared with that of $K^{\mathrm{P-NS}}$ (compare P–NS with app:P–NS) whereas the slopes of $K_{\mathrm{app}}^{\mathrm{P-S}}$ and $K^{\mathrm{P-NS}}$ are virtually identical (compare P–S with app:P–S).

In summary, the competitive interactions considerably reduce the values of the apparent affinity constants for specific and non-specific hybridization where the latter constant and thus also the non-specific background become partly insensitive for a particular probe sequence whereas the affinity for specific hybridization and thus also the concentration dependence depend strongly on the base composition of the probe.

3.8. Quality of the probes: specificity and sensitivity

The sensitivity of a probe characterizes its potential detection strength for specific targets under ideal conditions, i.e. in the absence of non-specific RNA fragments and in the absence of saturation. In contrast, the specificity of a probe characterizes its selectivity, i.e. its power for deciding between specific target RNA and the chemical background of non-specific RNA fragments. The hybridization isotherms of the DNA probes and the associated hybridization constants provide a natural starting point for the characterization of their quality as reporters for the concentration of specific target RNA in a complex mixture of RNA fragments in terms of sensitivity and specificity.

The apparent constant for specific hybridization, $K_{\mathrm{app}}^{\mathrm{P-S}}$, is directly related to the surface coverage due to specific transcripts at a given concentration of specific RNA. It therefore can serve as a measure of the sensitivity of a particular microarray probe. The apparent constant for

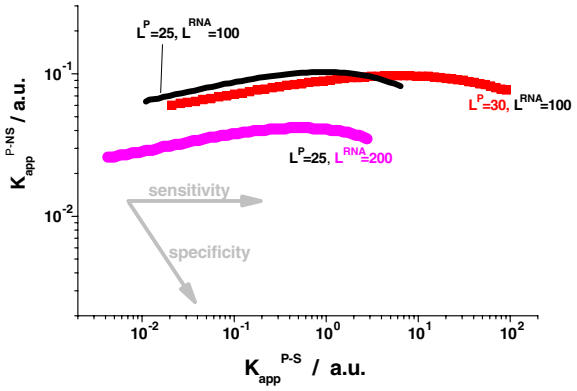

Figure 5. Correlation between the apparent association constants of non-specific and specific hybridization for probes and RNA fragments of different lengths (see also figure 3). The quality of the probes in terms of sensitivity and specificity increases in the direction of the arrows. The curves relate to the change of the GC content, from $x_{GC} = 0$ to 1.

non-specific hybridization, K_{app}^{P-NS}, is directly related to surface coverage due to non-specific transcripts. The ratio, $r_{app} \equiv K_{app}^{P-NS}/K_{app}^{P-S}$ consequently characterizes the relative amount of non-specific hybridization and thus it is inversely related to specificity of the corresponding probe.

Figure 5 correlates the apparent hybridization constants for non-specific and specific hybridization upon varying GC content. The arrows point in the direction of increased quality of the probes. The lengthening of the probe from $L^P = 25$ to 30, for example, improves its sensitivity and, but to a lesser degree, its specificity as well. This trend agrees with the results of hybridization studies with probes of varying length [24, 50].

On the other hand, the presence of mismatches in the probe sequence with respect to the specific target effectively reduces the probe length by the number of mismatched base pairings and thus also the value of K_{app}^{P-S}. On the other hand, K_{app}^{P-NS} remains unchanged because the 'mismatches' relate to the specific transcripts and not to the non-specific ones. The quality of mismatched probes is consequently smaller than that of perfectly matched ones. This trivial result reflects the reduced binding strength of mismatched probes [20].

Finally, we estimated the effect of longer RNA fragments which reduce both the apparent binding constants for specific and non-specific hybridization, $K_{app}^{P-h}(h = S, NS)$, due to the higher propensity of the RNA fragments for intramolecular folding and bulk dimerization (see the line for $L^{RNA} = 200$ in figure 5). Hence, longer RNA fragments degrade the performance of the oligonucleotide probes in agreement with experimental studies which found that short target lengths facilitate probe binding [24].

Here we considered the effect of only a few parameters on the probe quality for illustration. The selection of optimal probes requires a more comprehensive analysis using additional criteria such as the genome-wide uniqueness of the probe sequence and, in addition, appropriate experimental validation strategies [51, 52].

4. Electrostatic and entropic blocking

Experiments on DNA arrays have revealed a considerable decrease of the thermodynamic stability of surface tethered DNA/RNA hybrid duplexes compared with free duplexes in solution [19, 21]. The effect increases as the surface density of probes increases. These trends

were explained by the so-called blockage of the hybridization reaction, i.e. the progressive hampering of duplex formation owing to electrostatic and entropic repulsion between the assayed RNA targets in solution and the DNA probes on the chip surface [22, 34].

4.1. Surface electrostatics upon microarray adsorption

The effect of electrostatics near surfaces can be understood in terms of the surface partition model. It assumes that the formation of the probe/target duplex is governed by the apparent association constant, $K_{\mathrm{app}}^{\mathrm{P-h}}$ (h = S, NS), which characterizes the effective chemical binding affinity in the absence of electrostatics according to reaction scheme 1 (see figure 1). The electrostatic repulsion between the free RNA in solution on one hand and the surface grafted probe oligomers and always bound RNA on the other hand depletes the concentration of the dissolved RNA in the vicinity of the surface proportionally to the *Boltzmann* factor,

$$[\mathrm{h}]_{\mathrm{surface}} \approx [\mathrm{h}] \cdot \exp(-G_{\mathrm{el}}/RT) \qquad (\mathrm{h} = \mathrm{S}, \mathrm{NS}), \tag{4.1}$$

where $G_{\mathrm{el}} \approx F \cdot L^{\mathrm{RNA}} \cdot q_{\mathrm{N}} \cdot \psi_{\mathrm{s}}$ is the electrostatic free energy of a charged RNA fragment of length L^{RNA} within the surface potential ψ_{s} (F is the Faraday constant). The effective charge per nucleotide, q_{N}, takes into account electrostatic screening and thus it depends on the salt concentration and the permittivity of the aqueous medium (see, e.g., [53]).

The electrostatic potential is produced by the charged surface of the chip. Its two-dimensional charge density increases linearly with the surface coverage, θ, i.e.,

$$\sigma_q = \rho_{\mathrm{P}} \cdot L^{\mathrm{P}} \cdot q_N(1 + \theta \cdot r_L). \tag{4.2}$$

Here, ρ_{P} is the two-dimensional density of the oligonucleotides grafted on the chip. L^{P} and r_L denote the probe length and the ratio $r_L = L^{\mathrm{RNA}}/L^{\mathrm{P}}$. Note that the length of the RNA fragments typically exceeds the length of the probe, i.e. $r_L > 1$ (see above).

The surface potential and thus the electrostatic free energy can be calculated as a function of the surface charge (equation (4.2)) using the Grahame equation for 1:1 electrolytes, $\psi_{\mathrm{s}} \sim -(RT/F) \cdot \mathrm{arcosh}((\sigma_q)^2/4000\varepsilon_{\mathrm{W}}C_{\mathrm{NaCl}} + 1)$ (ε_{W} is the permittivity of water and C_{NaCl} the salt concentration). The binding rate at the surface is directly related to the concentration of the free adsorbate in an infinitely thin layer of the solution directly above the surface. The binding strength is consequently given by $X_{\mathrm{el}}^{\mathrm{h}} = K_{\mathrm{app}}^{\mathrm{P-h}} \cdot [\mathrm{h}]_{\mathrm{surface}}$ instead of $X^{\mathrm{h}} = K_{\mathrm{app}}^{\mathrm{P-h}} \cdot [\mathrm{h}]$. Insertion of equation (4.1) provides the adsorption strength as a linear function of the bulk concentration, i.e.,

$$X_{\mathrm{el}}^{\mathrm{h}} = K_{\mathrm{el}}^{\mathrm{h}} \cdot [h] \qquad \text{with } K_{\mathrm{el}}^{\mathrm{h}} \approx K^{\mathrm{h}} \cdot \exp(-G_{\mathrm{el}}(\theta)/RT), \qquad \mathrm{h} = \mathrm{S}, \mathrm{NS}. \tag{4.3}$$

The electrostatic free energy decomposes into a initial contribution due to the electrostatic repulsion produced by the free probes, G_{el}^0, and into an incremental term caused by the bound RNA which increases with the surface coverage, $\Delta G_{\mathrm{el}}(\theta)$, according to equation (4.2), $G_{\mathrm{el}}(\theta) = G_{\mathrm{el}}^0 + \Delta G_{\mathrm{el}}(\theta) \propto \sigma_q \propto \rho_{\mathrm{P}} \cdot L^{\mathrm{P}} \cdot (1 + \theta \cdot r_L)$. The overall hybridization constant consequently becomes a function of the surface coverage θ. It can be written as the product of nested constants

$$K_{\mathrm{el}}^{\mathrm{h}} = K^{\mathrm{h}} \cdot K_{\mathrm{el}} \cdot \delta K_{\mathrm{el}}(\theta), \tag{4.4}$$

which relate to chemical binding (K^{h}), initial and incremental electrostatic blocking, $K_{\mathrm{el}} = \exp(-G_{\mathrm{el}}^0/RT)$ and $\delta K_{\mathrm{el}}(\theta) = \exp(-\Delta G_{\mathrm{el}}(\theta)/RT)$, respectively. The factor $1/K_{\mathrm{el}}$ specifies the shift of the *Langmuir* hybridization isotherm along the concentration axis whereas $\delta K_{\mathrm{el}}(\theta)$ affects its slope and thus the deviation from the *Langmuir*-type behaviour. Note that $\Delta G_{\mathrm{el}}(\theta)$ exceeds G_{el}^0 at full surface coverage, $\theta = 1$, by the factor r_L (see equation (4.2)). The condition $\Delta G_{\mathrm{el}}(\theta) = G_{\mathrm{el}}^0$ provides the critical surface coverage $\theta_{\mathrm{el}} = 1/r_L$ at which the incremental factor becomes the leading contribution in equation (4.4) for $\theta > \theta_{\mathrm{el}}$.

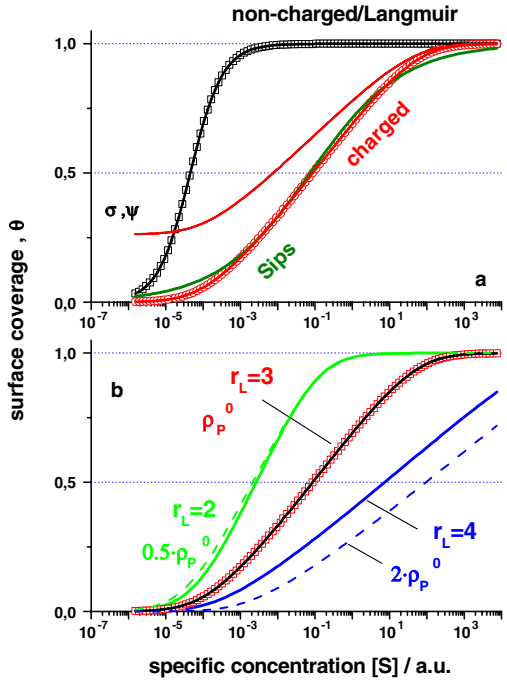

Figure 6. Electrostatic blocking of the hybridization efficiency of surface grafted probes. The isotherm which considers surface electrostatics (part (a): circles, 'charged'; see equations (4.1)–(4.3); the surface density of probes is $\rho_P^0 = 1/10^3$ nm^2, their GC content $x_{GC} = 0.5$ and the length ratio, target to probe $r_L = 3$) is considerably shifted towards the right and increases more slowly with [S], the concentration of specific transcripts, compared with the corresponding *Langmuir* isotherm which neglects surface electrostatics (squares, 'non-charged', [NS] = 0). The curves labelled σ, ψ show the corresponding 'electric' coverage, θ_σ and θ_ψ. The two curves are not distinguishable in the figure and, in turn, strongly correlate with material coverage, θ. The *Sips* isotherm ('Sips'; see equation (8.4)) approximates well the 'charged' isotherm using a Sips exponent of $a = 0.35$. Part (b): electrostatic blocking increases with increasing surface density of probes (dashed curves), and with increasing relative length of the RNA fragments (solid curves; see the figure for labels). Both trends shift the isotherms to the right and decrease their slope.

Accordingly, the depletion of the adsorbate near the surface gives rise to a progressively decreased binding rate compared with the corresponding reaction in the bulk solution. In other words, electrostatic blocking multiplicatively reduces the hybridization constants by the corresponding *Boltzmann* factor in a concentration dependent manner in addition to the competitive interactions discussed in the previous sections.

4.2. Sorption isotherms: the effect of DNA probe density and of RNA fragment length

Equations (4.1)–(4.3) were numerically solved to provide the surface coverage in equations (2.9) and (4.3) as a function of specific transcript concentration [S] in the absence of cross-hybridization ([NS] = 0). The consideration of electrostatic interactions markedly modifies the sorption isotherm with respect to the concentration at half-saturation, $[S]_{50\%}$, and with respect to its slope in the sigmoidal range at $[S] = [S]_{50\%}$ (see figure 6, part (a)). On the one hand, electrostatic blocking effectively reduces the hybridization affinity as stated above and thus $[S]_{50\%}$ shifts to bigger concentration values. On the other hand, electrostatics broadens the concentration range of sorption and thus the slope of the sorption isotherm decreases with

respect to the logarithmic concentration axis. The increase of the surface density of attached probes and/or of the length of the RNA fragments further amplify these trends (see part (b) of figure 6). The latter result shows that shorter RNA fragments bind with higher affinity than longer ones owing to the weaker electrostatic repulsion with the chip surface. One can conclude that the surface binding reaction preferentially selects relatively short target fragments from the available cocktail of different fragment lengths.

Note the 1:1 agreement between the change of the surface coverage, θ, and the electric 'coverage' of the surface calculated as $\theta_e = e/e_{\max}$ with $e = \psi, \sigma_q$ (see part (a) of figure 6; $e_{\max}$ is the asymptotic value at large [S]). It shows that the adsorption process is strongly modulated by surface electrostatics. Our simple approach neglects the thickness of the adsorption layer in the direction perpendicular to the microarray surface. For a vertically extended surface layer of thickness $\sim L^{\mathrm{P}}$, a more sophisticated analysis provides $\psi_{\mathrm{s}} \propto \sigma_q/L^{\mathrm{P}}$ [22] instead of $\psi_{\mathrm{s}} \propto \sigma_q$. Hence, the amount of electrostatic blocking is simply scaled by a thickness factor without qualitative consequences for sorption isotherms discussed here.

The surface density of the probe oligomers, the length distribution of RNA fragments and the thickness of the probe layer are common to all probes of a chip in a first-order approximation. Moreover, the variation of the affinity of a probe, for example on changing their GC content, shifts the Langmuir isotherm (without considering electrostatics) and the corresponding isotherm which considers electrostatic blocking by an identical increment along the concentration axis (not shown). Hence, electrostatics is expected to affect the isotherms of all probes of the chip in a unique, probe-unspecific manner with respect to their slope and position.

4.3. Entropic blocking

Single-stranded DNA and RNA fragments are flexible polymeric chains. The free oligomeric DNA probes and the RNA fragments which are bound to the probes are fixed to the chip surface and thus they possess the characteristics of surface grafted polymers. Essentially two factors can reduce their conformational freedom compared with free polymers in solution. Firstly, the impermeable solid surface which fixes one end of the polymeric chain and, secondly, the crowding with neighbouring polymers the tails of which overlap with the region accessible by the polymer considered. These phenomena were extensively studied in polymer physics and they apply also to the hybridization reaction on microarrays as has been shown recently (see [54] and references cited therein). Accordingly, the free DNA probes and probes with bound RNA form a 'turf' of grafted oligomers. It produces a repulsive, entropic force on the free RNA fragments in the supernatant solution due to the reduction of conformational freedom upon transfer of the RNA from bulk solution into the bound, i.e., grafted state. The entropic penalty hampers the binding of free RNA to the free probes as a consequence.

The probe bound RNA fragments are typically more important in the context of entropic blocking than the free probes because the length of the former species usually exceeds that of the latter ones, i.e., $L^{\mathrm{RNA}} > L^{\mathrm{P}}$ (see also below). Essentially one has to consider three situations [54] (see figure 7 for illustration). (a) The '1:1' mushroom regime where bound RNA interacts essentially only with the DNA probe which 'chemically' binds the RNA in a 1:1 fashion. This regime applies to low probe densities and small surface coverage if a bound RNA fragment is predominantly surrounded by free probes. It does not interact with the neighbouring free probes because the characteristic hemisphere radius of the bound RNA, R_{F}, is smaller than the distance between adjacent probes. The '1:1' regime applies consequently to small surface densities of the probes, $\rho_{\mathrm{P}} < R_{\mathrm{F}}^{-2}$ which are characterized by the critical surface coverage of $\theta_{\mathrm{B}} = \min(1, 1/(\rho_{\mathrm{P}} \cdot R_{\mathrm{F}}^2)) = 1$. (b) The '1:$q$' mushroom regime relates to $\theta < \theta_{\mathrm{B}} < 1$ where bound RNA interacts with $q \approx 1/\theta_{\mathrm{B}} > 1$ neighbouring free probes but not with other bound

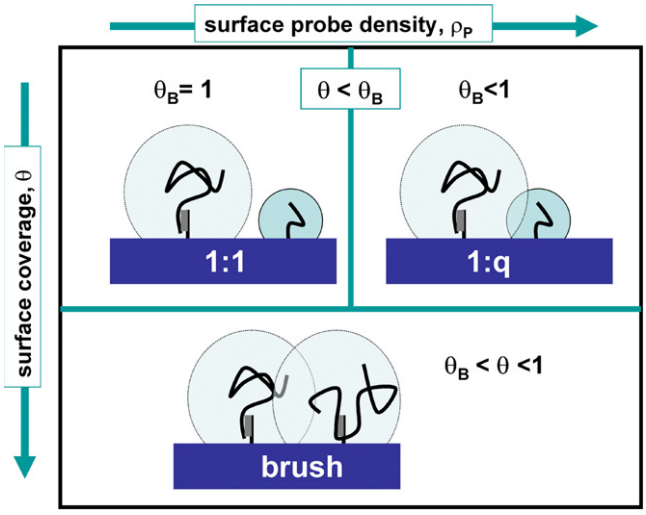

Figure 7. Entropic blocking of surface grafted oligomer probes. The figure schematically illustrates the 1:1 and 1:q mushroom and the polymeric brush regimes. The hemispheres illustrate the region which is filled by the conformations of the polymeric tails. The region of overlap between neighbouring spheres consequently restricts their conformational freedom. This entropic penalty gives rise to the blocking of the hybridization reaction. The smaller spheres relate to the free probes whereas the larger ones refer to probe bound RNA fragments.

RNA fragments. This regime consequently applies to higher probe densities, $\rho_P \approx R_F^{-2}$, and small surface coverage, $\theta < \theta_B$. (c) In contrast, the polymer brush regime applies to higher surface coverage which exceed the critical value, $\theta \geqslant \theta_B$, if RNA fragments bound to adjacent probe oligomers significantly interact with each other.

The entropy penalty of the '1:1' mushroom regime is caused by the impermeable surface and thus it is independent of the surface coverage. The entropy penalty in the '1:q' mushroom regime in addition depends on the conformational restrictions owing to neighbouring probes but it remains independent of θ. In contrast, the entropy penalty in the polymer brush regime strongly increases with θ because the conformational freedom of bound RNA fragments decreases if adjacent sites become progressively occupied. The entropic repulsion decreases the free energy gain upon duplex formation and thus it effectively decreases the association constant which can be written in analogy with equation (4.4) as the product of nested constants

$$K_S^h = K^h \cdot K_S \cdot \delta K_S(\theta). \tag{4.5}$$

A microscopic model [54] provides

$$\begin{aligned} K_S &\approx (r_L)^{-2/5} \cdot \exp\left(\frac{(L^P)^{1/5}}{r_L^{4/5}} \cdot (1 - \theta_B^{-1})\right); \\ \delta K_S &\approx \begin{cases} 1 & \text{for } \theta < \theta_B \\ \theta^{*\frac{1}{3}} \cdot \exp(-(L^P \cdot (r_L - 1))^{1/5} \cdot (\theta^{*\frac{2}{3}} - 1)) & \text{for } \theta \geqslant \theta_B \end{cases} \\ \text{with } \theta^* &\equiv \frac{\theta}{\theta_B}, \quad \theta_B \equiv \frac{1}{\rho_P \cdot R_F^2} \quad \text{and} \quad R_F = l_{\text{segment}} \cdot (L^P \cdot (r_L - 1))^{3/5}. \end{aligned} \tag{4.6}$$

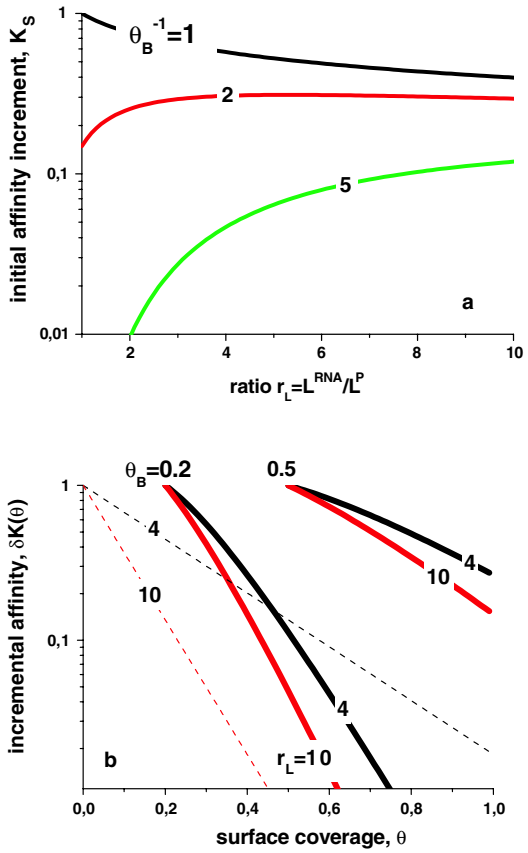

Figure 8. Entropic blocking of the hybridization efficiency of surface grafted probes: part (a) shows the initial affinity increment as a function of the relative length of the RNA fragments in the 1:1 and 1:q(q = 2, 5) mushroom regime (see equation (4.6)). Part (b) shows the decrease of the incremental binding affinity with increasing surface coverage for two ratios $r_L = L^{\mathrm{RNA}}/L^{\mathrm{P}}$ and two values of the critical surface coverage (see figure and equation (4.6)). The dotted lines relate to the incremental affinity due to electrostatic blocking.

The '1:1' and '1:q' mushroom regimes relate to $\theta_{\mathrm{B}} = 1$ and $\theta < \theta_{\mathrm{B}} < 1$, respectively. The hybridization in the mushroom regimes follows the *Langmuir* isotherm the $[\mathrm{S}]_{50\%}$ value of which is however increased by the factor $1/K_{\mathrm{S}}$ due to entropic blocking. Part (a) of figure 8 shows K_{S} as a function of the relative length of the RNA fragments, r_L. The initial constant of entropic blocking decreases with increasing number of lateral interaction sites, $q \approx \theta_{\mathrm{B}}^{-1}$ because of the progressive overlap of the tails of bound transcript with neighbouring oligonucleotide probes. Entropic blocking can reduce the hybridization constant by more than one order of magnitude even in the mushroom regime.

At $\theta = \theta_{\mathrm{B}}$ the '1:q' mushroom regime crosses over into the polymer brush regime. It is characterized by the progressive decrease of δK_{S} with increasing surface coverage for $\theta > \theta_{\mathrm{B}}$. The corresponding δK_{S} term in equation (4.6) considers the fact that only the flexible single-stranded dangling ends of the RNA contribute to the entropic blocking in the duplexes with the probes because the double-stranded region acts rather as a stiff rod than a flexible polymeric tail. Part (b) of figure 8 shows the decrease of the incremental affinity constant with increasing

surface coverage. The effect becomes stronger for longer fragments and/or a smaller critical surface coverage, θ_B, because both effects amplify entropic blocking.

The dashed lines in part (b) of figure 8 show the incremental term of electrostatic blocking, $\delta K_{el}(\theta)$, for comparison with $\delta K_S(\theta)$. The two blocking mechanisms obviously give rise to similar trends upon increasing surface coverage. The entropic and electrostatic blocking mechanisms modify the isotherm in similar fashions (see figure 6 and [54]). Moreover, as in the case of electrostatics, entropic blocking is expected to affect all probes identically because the surface density, probe length and average RNA fragment length are common to all probes.

Electrostatic and entropic blocking mechanisms essentially act independently of each other. In this case equations (4.4) and (4.5) merge into

$$\begin{aligned} &K_B^h = K^h \cdot K_B \cdot \delta K_B(\theta) \\ &\text{with } K_B \equiv K_{el} \cdot K_S \quad \text{and} \quad \delta K_B(\theta) \equiv \delta K_{el}(\theta) \cdot \delta K_S(\theta). \end{aligned} \tag{4.7}$$

The effects of electrostatic and entropic blocking on the $[S]_{50\%}$ value amplify each other in a multiplicative fashion.

5. Fragmentation of the RNA transcripts and truncation of the DNA probes

In the previous sections we assumed a homogeneous ensemble of 'ideal' probe–target duplexes which associate via L^P complementary WC base pairings over the whole 'nominal' length of the probes, $L^P = L^P_{nom}$ (for example with $L^P_{nom} = 25$ for GeneChip oligomers). In this section we discuss the effect of the truncation of both target and probe, due to the fragmentation of the RNA and due to imperfect photolithographic synthesis of the oligomer probes on the chip. Both effects destabilize the hybrid duplexes on a relative scale owing to the reduced number of WC pairings.

5.1. Fragmentation of the RNA transcripts

In the fragmentation step the cRNA is randomly cut into pieces with an average length L^{RNA}. Fragments which are cut inside the target region provide a reduced number of complementary 'target' bases for duplex formation with the probe (i.e., $L < L^P$), compared with fragments possessing an 'intact' target region with $L = L^P$ complementary bases (see figure 2). The fraction of specific RNA fragments with L complementary matches in the sample solution is (see also equation (3.5))

$$P_{frag}(L, L^P) = \begin{cases} 2/(L^{RNA} + L^P - 1) & \text{for } L = 1, \ldots, L^P - 1 \\ N_{match}(L, L^{RNA})/(L^{RNA} + L^P - 1) & \text{for } L = L^P. \end{cases} \tag{5.1}$$

For simplicity we assume a uniform length of the fragmented RNA of $L^{RNA} = \text{constant} = 100$. It turns out that only about 60% of all fragments completely match the full probe length of $L^P = 25$. The corresponding average number of matched bases, $\langle L \rangle = \sum L \cdot P_{frag}(L, 25) \approx 20$, is considerably smaller than the probe length. Upon hybridization the different target fragments compete with each other for formation of duplexes with the probe oligonucleotides. The effective association constant of specific hybridization is given as the weighted mean over the available target lengths (compare with equation (3.3))

$$K^{P-S}_{frag}(L^P) = \sum_{L=1}^{L^P} P_{frag}(L, L^P) \cdot K^{P-S}(L) = K^{P-S}(L^P) \cdot K_{frag}. \tag{5.2}$$

The resulting association constant is smaller than $K_0^{P-S}(L^P)$ by the factor $K_{frag} < 1$ owing to the truncated RNA fragments. The free energy function used in the previous section

(equations (3.2) and (3.3)) provides a ratio $K^{\mathrm{P-S}}_{\mathrm{frag}}(25)/K^{\mathrm{P-S}}(25) \approx 0.69$–$0.63$ for $x_{\mathrm{GC}} = 0.0$–1.0 and $L^{\mathrm{RNA}} = 100$, i.e. a moderate decrease of the binding constant due to RNA fragmentation. The effect of truncation of the target region decreases with increasing length of the fragments. For example, one obtains $K^{\mathrm{P-S}}_{\mathrm{frag}}(25)/K^{\mathrm{P-S}}(25) \approx 0.96$ for $L^{\mathrm{RNA}} = 1000$.

5.2. Incomplete synthesis of the oligonucleotide probes

Methods for microarray fabrication include spotting of DNA fragments of genomic DNA, cDNA, PCR products or chemically synthesized oligonucleotides onto nylon membranes or glass slides by robots with pins or inkjet printers [50, 55, 56]. cDNA arrays are often used in RNA expression analysis, while oligonucleotide arrays are additionally used for genomic sequence analyses using SNP or tiling arrays [10, 11, 57, 58]. The oligonucleotide arrays are fabricated either by conventional synthesis followed by immobilization on the substrate or by *in situ* light-directed combinatorial synthesis on the surface of the array. This photolithographic technique combines solid phase chemical synthesis with photolithographic fabrication employed in the semiconductor industry. It enables us to produce microarrays of very high density. Current state-of-the-art technology allows the inclusion of more than 10^6 sequences representing nearly 50 000 genes on a surface of 1.6 cm^2.

The photolithographic synthesis proceeds step by step. In particular, the protective group which finalizes the oligomers is selectively removed by light exposure using a lithographic mask which spatially selects the corresponding probes on the chip. Then, the hydroxyl protected nucleoside which relates to the next sequence position (A, T, G or C) is coupled to the deprotected end of oligomer. This process is repeated $L^{\mathrm{P}}_{\mathrm{nom}}$ times to synthesize oligomers of nominal length $L^{\mathrm{P}}_{\mathrm{nom}}$. The length of all oligomers of one probe ideally grows by one nucleotide in each step.

The efficiency of deprotection is however the limiting factor in this procedure, reported to result in overall stepwise synthetic yields in the range of 82–98% [59–61]. Oligonucleotides that fail to be photodeprotected as intended are irreversibly blocked and cannot be extended in subsequent cycles, yielding 5′ truncated products but not internal deletions.

Hence, a few per cent of all oligomers become finalized after each step and thus the length of these oligomers is smaller than the nominal one, i.e., $L^{\mathrm{P}} < L^{\mathrm{P}}_{\mathrm{nom}}$. The fraction of oligomers of length L^{P} is

$$P_{\mathrm{synt}}(L^{\mathrm{P}}, L^{\mathrm{P}}_{\mathrm{nom}}) = \begin{cases} p_{\mathrm{synt}}^{L-1}(1 - p_{\mathrm{synt}}) & \text{for } L^{\mathrm{P}} = 1, \ldots L^{\mathrm{P}}_{\mathrm{nom}} - 1 \\ p_{\mathrm{synt}}^{L-1} & \text{for } L^{\mathrm{P}} = L^{\mathrm{P}}_{\mathrm{nom}}. \end{cases} \tag{5.3}$$

For example, a synthesis yield of $p_{\mathrm{synt}} \approx 0.90$–$0.98$ provides a fraction of only $P_{\mathrm{synt}}(25, 25) \approx 0.1$–$0.6$ for oligomers of nominal length $L^{\mathrm{P}}_{\mathrm{nom}} = 25$ in the corresponding probe spot. The mean length of the oligonucleotides reduces considerably to $\langle L^{\mathrm{P}} \rangle = \sum L \cdot P(L, 25) \approx 9$–$20$ compared with $L^{\mathrm{P}}_{\mathrm{nom}} = 25$. Hence, the *in situ* synthesis procedure results in arrays in which the oligonucleotide features are heavily contaminated with truncated versions of the desired probe sequences.

5.3. Heterogeneous adsorption

The oligonucleotide probes of different length form a heterogeneous ensemble where each fraction of length L^{P} serves as an independent adsorption site. The superposition of their

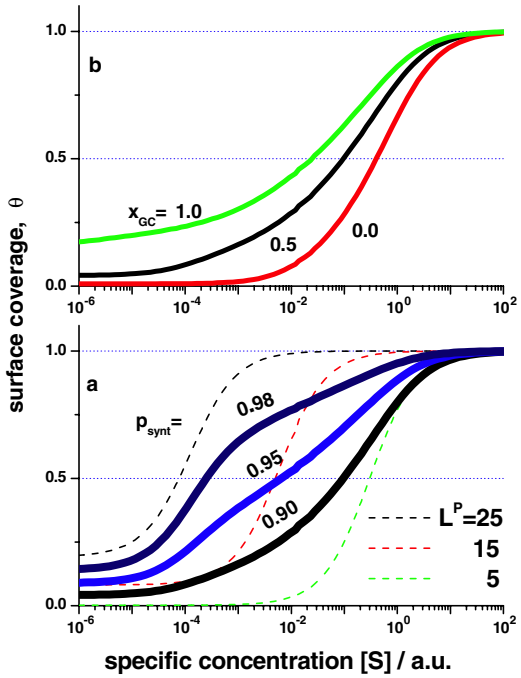

Figure 9. Heterogeneous adsorption owing to truncated probes after incomplete synthesis (see equations (5.3) and (5.4)). The isotherms shift to the right with decreasing synthesis yield, p_{synt} (part (a), thick curves, $x_{\mathrm{GC}} = 0.5$). Note that the fraction of probes of full length ($L^{\mathrm{P}}_{\mathrm{max}} = 25$) decreases from 0.6 to 0.1 if p_{synt} decreases from 0.98 to 0.90. The dashed curves are the corresponding *Langmuir* isotherms for truncated probes of length L^{P} (see the figure for labels). Part (b) shows that the isotherms are affected by the base composition of the probes. Upon varying the GC content heterogeneous adsorption changes the position of the inflection point, the slope and the background coverage at small [S]. Note that the stability of the duplexes increases with the GC content according to our energy function (see the legend of figure 3 and equation (3.2)).

coverage provides the overall hybridization isotherm

$$\theta = \sum_{L^{\mathrm{P}}=1}^{L^{\mathrm{P}}_{\mathrm{nom}}} P_{\mathrm{synt}}(L^{\mathrm{P}}, L^{\mathrm{P}}_{\mathrm{nom}}) \cdot \theta(L^{\mathrm{P}}) \qquad \text{with } \theta(L^{\mathrm{P}}) = X(L^{\mathrm{P}})/(1 + X(L^{\mathrm{P}})) \tag{5.4}$$

where the binding strength, $X(L^{\mathrm{P}})$, splits into the contributions of specific and non-specific binding according to equation (2.9). Figure 9 illustrates the effect of heterogeneous adsorption on the binding isotherms as a function of specific transcript concentration, [S], using the microscopic model to calculate the binding constants of specific and non-specific hybridization as a function of the probe length L^{P} (see equation (3.1)).

The 'individual' adsorption isotherms of probes of length L^{P}, $\theta(L^{\mathrm{P}})$, show a Langmuir-type behaviour. Their inflection point, $[\mathrm{S}]_{50\%} \sim 1/K^{\mathrm{P-S}}(L^{\mathrm{P}})$, shifts towards larger concentrations of specific transcripts with decreasing probe length owing to the decrease of $K^{\mathrm{P-S}}(L^{\mathrm{P}})$ (compare the dashed curves in figure 9, part (a)). This tendency is paralleled by the decrease of the background level of non-specific hybridization which becomes apparent at small $[\mathrm{S}] \ll [\mathrm{S}]_{50\%}$. Note that also the association constant for non-specific hybridization decreases with L^{P} (see above) and in this way reduces the background level owing to the relation $\theta(L^{\mathrm{P}})|_{[\mathrm{S}] \ll [\mathrm{S}]50\%} \propto K^{\mathrm{P-NS}}(L^{\mathrm{P}})$.

The horizontal shift of the individual isotherms with respect to each other reflects the fact that the longer oligomers effectively hybridize at smaller target concentrations than the shorter ones due to their stronger affinity. In other words, the oligonucleotides of different length hybridize 'consecutively' with increasing [S] in the direction from longer to shorter oligomers. The overall adsorption isotherm represents the weighted mean of the individual curves relating to different probe lengths (see equation (5.4)). Its shape therefore resembles that of the individual isotherms of longer probes at small concentration of specific transcripts and that of shorter probes at larger [S] (see part (a) of figure 9). The overall isotherm of the ensemble of truncated probes is shifted towards higher concentrations compared with the 'ideal' isotherm relating to the homogeneous ensemble of probes of nominal length indicating the better performance of the latter probes. It was indeed experimentally verified that the specificity of pure oligonucleotide probes of one length was significantly greater than that of a population of an ensemble of truncated probes [62].

The increase of the concentration of half-coverage due to the truncation of the probes can be formally considered by the reduction factor $K_{\mathrm{trunc}} < 1$,

$$K_{\mathrm{trunc}}^{\mathrm{P-h}} = K^{\mathrm{P-h}} \cdot K_{\mathrm{trunc}} \approx 1/[\mathrm{S}]_{50\%}. \tag{5.5}$$

The fraction of longer oligomers, especially that of the nominal probe length, distinctly increases with increasing synthesis yield, p_{synt} (see equation (5.3)), with consequences for the background level, the slope and the inflection point of the isotherms (compare the thick curves in part (a) of figure 9). Also the GC content affects the shape of the overall adsorption isotherm (see part (b) of figure 9). In a more general context this result leads to the conclusion that the shape of the overall isotherms depends in a specific fashion on the sequence of a particular oligomer probe.

6. Zipping of hybrid duplexes

6.1. Specific duplexes

The kinetics of the formation of a duplex between the probe and target can be split into two steps. (i) The probe 'captures' the target by forming the first few base pairings of the duplex (nucleation reaction) after diffusion of the target into the vicinity of the probe. The rate of this step is mainly affected by the target concentration. (ii) The 'zipping' of the duplex in the direction towards the ends of the probe, i.e., the successive 'closure' of bonds between complementary bases. Vice versa, also always formed duplexes which are stabilized via WC pairings along the whole sequence can 'unzip' partially or even completely. Hence, zipping and unzipping of bonds should be analysed using equilibrium statistical mechanics. This approach was recently applied to calculate the hybridization isotherms of GeneChip probes [49]. It has been assumed that the relatively large local stiffness of double-stranded DNA/RNA hybrids prevents unzipping in isolated islands in the middle of the duplexes. This assumption is supported by the fact that short duplexes usually melt through end openings [63]. Therefore only configurations are considered for which the unbound parts start at one or both ends of the duplex.

Accordingly, the association constant of specific hybridization is given by the partition function of all zipped microstates with $1 \leqslant L \leqslant L^{\mathrm{P}}$ closed base pairings (see also equations (3.1) and (3.5)),

$$P_{\mathrm{match}}(L, L^{\mathrm{P}}) \equiv N_{\mathrm{match}}(L, L^{\mathrm{P}}) \Big/ \sum_{L=1}^{L^{\mathrm{P}}} N_{\mathrm{match}}(L, L^{\mathrm{P}}). \tag{6.1}$$

The consideration of partly zipped microstates decreases the value of the binding constant compared with $K^{\mathrm{P-S}}(L^{\mathrm{P}})$, the binding constant of the most stable duplex with L^{P} base pairings.

A potential base pairing is defined as 'open' if it does not contribute to the free energy of the system. The probability that a base pair at position $1 \leqslant k \leqslant L^{\mathrm{P}}$ remains unzipped is consequently

$$P_{\mathrm{open}}(k) = \frac{K^{\mathrm{P-S}}_{\mathrm{open}}(k)}{K^{\mathrm{P-S}}_{\mathrm{zipp}}} \tag{6.2}$$

where the denominator defines the full partition function of all possible microstates (equation (6.1)). The numerator considers all states with the open base pairing at position k. It can be split into two terms which relate to microstates with closed pairings at $1 \leqslant L < k$ and $k < L \leqslant L^{\mathrm{P}}$, respectively:

$$K^{\mathrm{P-S}}_{\mathrm{open}}(k) = K^{-}_{\mathrm{open}}(k) + K^{+}_{\mathrm{open}}(k)$$

with

$$K^{-}_{\mathrm{open}}(k) = \sum_{L=1}^{k-1} P_{\mathrm{match}}(L, k-1) \cdot K_0^{\mathrm{P-S}}(L) \qquad \text{and}$$

$$K^{+}_{\mathrm{open}}(k) = \sum_{L=k+1}^{L^{\mathrm{P}}} P_{\mathrm{match}}(L, L^{\mathrm{P}} - k + 1) \cdot K_0^{\mathrm{P-S}}(L).$$

Figure 10 (part (a)) shows the probability of 'closed' base pairings at position k, $P_{\mathrm{pair}}(k) = 1 - P_{\mathrm{open}}(k)$, which was calculated by means of the free energy function (equation (3.2)). The probability of paired bases decreases symmetrically towards the ends of the probe. A 25-meric poly-C probe ($x_{\mathrm{GC}} = 1$) is distinctly more 'closed' than a 25-meric poly-A oligomer ($x_{\mathrm{GC}} = 0$) because of the smaller contribution of a $A \bullet u$ pairing to the free energy of the duplex ($|\Delta g_{\mathrm{GC}}| > |\Delta g_{\mathrm{AT}}|$). Note that one GC base pairing is more stable than an AT pairing by $|\Delta g_{\mathrm{GC}} - \Delta g_{\mathrm{AT}}| = 0.4$ (in RT units) in the example shown. This relatively small free energy difference considerably modifies the probability distribution of closed base pairings because of the exponential relation between the free energy and the binding constant.

The free energy function used characterizes the stability of a duplex in terms of its mean base composition (equation (3.2)). A more sophisticated, sequence-specific description requires a position dependent consideration of, e.g., single-base related free energy terms according to

$$\begin{aligned} &G_0^{\mathrm{P-S}}(\xi^{\mathrm{P}}\xi^{\mathrm{S}})/RT = g_{\mathrm{init}} + \sum_{k=1}^{L^{\mathrm{P}}} g(\xi_k^{\mathrm{P}} \bullet \xi_k^{\mathrm{S}}) \\ &\text{with } g(\xi_k^{\mathrm{P}} \bullet \xi_k^{\mathrm{S}}) \approx g^{\mathrm{WC}}(\xi_k^{\mathrm{P}}) = g_0^{\mathrm{WC}} + \Delta g^{\mathrm{WC}}(\xi_k^{\mathrm{P}}). \end{aligned} \tag{6.3}$$

Here, ξ^{P} and ξ^{S} denote the sequences of the probe and the specific, complementary target, respectively. Within the dimer both the probe and target form the base pairing $\xi_k^{\mathrm{P}} \bullet \xi_k^{\mathrm{S}}$ at position k of the probe sequence (B = A, T, G, C and ξ_k^{S} = u*, a, c*, g; upper case letters relate to the DNA, lower case letters to the RNA, the asterisk indicates labelling). The approximation in the second line of equation (6.3) assumes that only WC pairings significantly contribute to the stability of the duplex. The $\Delta g^{\mathrm{WC}}(\mathrm{B})$ with $\mathrm{B} = \xi_k^{\mathrm{P}} = \mathrm{A, T, G, C}$ are the incremental, base-specific contributions of a single WC base pairing to the free energy of the duplex whereas g_0^{WC} denotes the mean free energy of a WC pair.

The total and the partial partition functions (equations (6.1) and (6.2)) can be calculated with equation (6.3) using the recursion relations (see [49] and references cited therein) $K^{+}_{\mathrm{open}}(k) = K^{+}_{\mathrm{open}}(k-1) + G^{+}(k-1)$ and $G^{+}(k) = G^{+}(k-1) \cdot \exp(g_0 + \Delta g(\xi_k^{\mathrm{P}})) + \exp(g_{\mathrm{init}})$

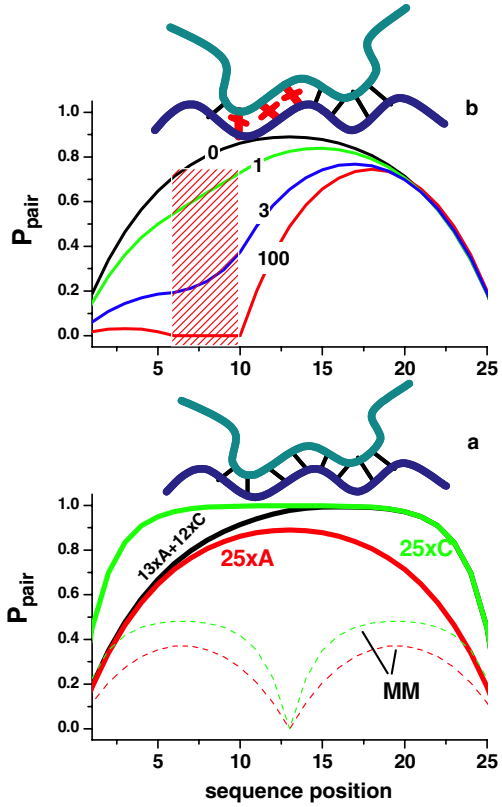

Figure 10. Zipping of DNA/RNA duplexes formed between a 25-meric poly-C (25 × C), a poly-A (25 × A) and a mixed composition 13 × A + 12 × C probe with the complementary target sequence (part (a)). The figure shows the probability of base pairings at each position of the P–S duplex. The dashed profiles relate to poly-A and poly-C probe sequences with a mismatched middle base which causes a strong free energy penalty relative to the corresponding WC pairing, $\delta g^{\mathrm{mm}} = 100$ (see the text). Part (b) shows the effect of varying the free energy penalty δg^{mm} for five mismatched bases at position 6–10 of a poly-A probe (see the figure for labels). The probability of paired bases decreases with increasing δg^{mm}. This trend leads to the progressive decoupling between the ranges of unfragmented WC pairings on both sides of the mismatches.

(for K^{-}_{open} substitute $k \to L^{\mathrm{P}} - k + 1$). Part (a) of figure 10 illustrates the trivial fact that the probability of paired bases becomes an asymmetrical function for asymmetrical sequences. For example, the specific duplex of a 13 × A + 12 × C probe zips with higher probability in the right, cytosine-rich part of the sequence due to the stronger base pairings.

6.2. Non-specific duplexes

The zipping of base pairings in the other complexes considered in figure 1 can be analogously described by means of the statistical approach. For example, a non-specific P–NS duplex is characterized by a certain number of mismatched base pairings along the probe sequence. Part (b) of figure 10 illustrates the effect of a mismatched region of five bases ranging from position $k = 6$–10 of a poly-A probe. Each mismatched pairing is characterized in equation (6.3) by a free energy contribution of g^{mm} instead of g^{WC}. The probability of paired bases at

the positions of the mismatches decreases with increasing free energy penalty per mismatch, $\delta g = -g^{\mathrm{WC-mm}}/g^{\mathrm{WC}} = -(g^{\mathrm{WC}} - g^{\mathrm{mm}})/g^{\mathrm{WC}}$.

More importantly, the presence of mismatches also decreases the probability of WC pairings in the complementary regions of the duplex. In the poly-A example shown in figure 10 the complementary bases at position $k = 1$–5 remain progressively unpaired with increasing energy penalty of the adjacent mismatches. This trend can be explained by the fact that the mismatches 'cut' the total number of $L = 20$ complementary bases into two fragments of length $L_1 = 5$ and $L_2 = 15$ which progressively decouple upon increasing δg due to the zipping effects.

This results shows that the assumption of additive free energy contributions of the total number of complementary pairings (see equations (3.1) and (3.2)) should be judged as one limiting case which estimates the upper limit of the corresponding binding constant, $K^{\mathrm{P-NS}}(L)$. It neglects the possible fragmentation of the total number of complementary bases L into M pieces of length L_m ($m = 1, \ldots, M$) separated by mismatches according to $L = \sum_m L_m$.

Equations (3.1) and (3.2) consequently relate to the limiting case of a relatively small free energy penalty of the mismatches, e.g., $\delta g^{\mathrm{mm}} \approx 1$ for $g^{\mathrm{mm}} \approx 0$, which provides $K_{>}^{\mathrm{P-NS}}(L) \approx K^{\mathrm{P-NS}}(L_1) \cdot K^{\mathrm{P-NS}}(L_2)$ for the chosen example. Another limiting case relates consequently to a high energy penalty per mismatch, e.g., $\delta g^{\mathrm{mm}} \approx 100 \gg 1$, which virtually splits the complementary region into almost independent fragments with adjacent WC pairings. This situation provides the lower limit of the binding constant, $K_{<}^{\mathrm{P-NS}}(L) \approx K^{\mathrm{P-NS}}(L_1) + K^{\mathrm{P-NS}}(L_2) \approx K^{\mathrm{P-NS}}(\max(L_1, L_2))$.

6.3. Isotherms

The zipping/unzipping scenario implies that the diffusion and nucleation step (i) is the rate limiting process which controls the hybridization reaction. In this case, the binding follows a Langmuir isotherm where the sorbate is bound by a homogeneous sorbent, i.e., the non-specific and specific RNA fragments hybridize the probes with effective binding constants which consider all zipped states (see equation (6.1) for specific hybridization).

7. Mismatched probes

Mismatched (MM) probes with a single non-complementary base pairing with respect to the specific target are designed in combination with the perfectly matched (PM) probes to estimate the amount of non-specific hybridization [13]. This pairwise combination of probes is based on the assumption that the affinity of the MM probes for specific transcripts is markedly reduced compared with that of the PM (i.e., $K^{\mathrm{PM-S}} \gg K^{\mathrm{MM-S}}$), whereas the affinities of the two kinds of probes for non-specific hybridization are virtually identical (i.e., $K^{\mathrm{PM-NS}} \approx K^{\mathrm{MM-NS}}$). The latter relation seems to be satisfied in a trivial fashion because the mismatched base is by definition related only to the specific target sequence and thus PM and MM are virtually equivalent with respect to a cocktail of different non-specific RNA fragments. In other words, $K^{\mathrm{PM-NS}}$ and $K^{\mathrm{MM-NS}}$ characterize complexes of the same type, the sequence of which effectively differs, however, by one position, namely the mismatched base (see [43]).

Also the stronger affinity of the PM for specific binding is plausible because the free energy penalty of the mismatched base pairing reduces the stability of the MM–S compared with that of the PM–S complex. The consideration of the two limiting cases of a small and a large energy penalty discussed in the previous section provides the corresponding binding constants of the MM, $K_{>}^{\mathrm{MM-S}}(\xi^{\mathrm{MM}}) \approx K^{\mathrm{WC}}(\xi_{L1}^{\mathrm{PM}}) \cdot K^{\mathrm{WC}}(\xi_{L2}^{\mathrm{PM}}) \cdot K^{\mathrm{mm}}(\xi_m^{\mathrm{MM}})$ and $K_{<}^{\mathrm{MM-S}}(\xi^{\mathrm{MM}}) \approx K^{\mathrm{WC}}(\xi_{L1}^{\mathrm{PM}}) + K^{\mathrm{WC}}(\xi_{L2}^{\mathrm{PM}})$, respectively. Here we take into account that the

MM–S duplex splits into two complementary regions of L1 bases before and L2 bases after the mismatch at position m. The contribution of the mismatch is $K^{\mathrm{mm}}(\xi_m^{\mathrm{MM}}) = \exp\{g(\xi_m^{\mathrm{MM}} \bullet \xi_m^{\mathrm{S}})\}$. The probability profiles of closed base pairings are illustrated for a mismatched poly-A and a poly-C probe in part (a) of figure 10.

The corresponding binding constant of the PM probe is $K^{\mathrm{PM-S}}(\xi^{\mathrm{PM}}) \approx K^{\mathrm{WC}}(\xi_{L1}^{\mathrm{PM}}) K^{\mathrm{WC}}(\xi_{L2}^{\mathrm{PM}}) K^{\mathrm{WC}}(\xi_m^{\mathrm{PM}})$ according to this notation. The effective loss of the affinity due to the presence of a single mismatch can be specified by the ratio of the binding constants for PM–S and MM–S duplexes, $K^{\mathrm{PM-MM,S}} \equiv K^{\mathrm{PM,S}}/K^{\mathrm{MM,S}}$ which becomes for both limiting situations considered

$$K_{>}^{\mathrm{PM-MM,S}} = \frac{K^{\mathrm{WC}}(\xi_m^{\mathrm{PM}})}{K^{\mathrm{mm}}(\xi_m^{\mathrm{MM}})} \approx \exp\{-(g^{\mathrm{WC}}(\xi_m^{\mathrm{PM}}) - g(\xi_m^{\mathrm{MM}} \bullet \xi_m^{\mathrm{S}}))\}$$
$$\text{and } K_{<}^{\mathrm{PM-MM,S}} \approx \frac{K^{\mathrm{WC}}(\xi_m^{\mathrm{PM}})}{K^{\mathrm{WC}}(\xi_{L1}^{\mathrm{PM}})^{-1} + K^{\mathrm{WC}}(\xi_{L2}^{\mathrm{PM}})^{-1}}. \tag{7.1}$$

The loss of affinity is exponentially related in an approximate fashion to the single-base free energy penalty due to the replacement of the WC pair at position m by the mismatched pair $\xi_m^{\mathrm{MM}} \bullet \xi_m^{\mathrm{S}}$ in the former case of small δg. A similar trend is predicted for large δg. The contribution of the WC pair at position m is however reduced by a factor which depends on the whole remaining sequence of the probe. As a consequence, a different behaviour of the affinity loss upon changing the position of the mismatch along the sequence is predicted: in the former case the free energy penalty is of the order of the mean free energy contribution of a WC pair independently of the sequence position. Hence, $K_{>}^{\mathrm{PM-MM,S}} \approx \exp(-g^{\mathrm{WC}}(\xi_m^{\mathrm{PM}})) \approx \exp(-g_0^{\mathrm{WC}})$ is almost invariant along the sequence if one neglects the base-specific incremental contribution to the free energy difference and zipping effects. In contrast, the second limiting situation predicts a varying affinity loss along the sequence, $K_{<}^{\mathrm{PM-MM,S}} \approx \exp(-(\min(L_1, L_2) + 1) \cdot g_0^{\mathrm{WC}})$, which is maximum for the mismatch position in the middle of the sequence and which is minimum for mismatched pairing in the first or the last position of the probe sequence. Hence, a mismatched base in the middle of the sequence is expected to reduce the stability of the duplex most effectively.

Experimental studies clearly reveal agreement with the latter prediction [64, 65]. On the other hand it was found that the effect of the mismatches remains virtually constant if its position is varied in the central part of the probe, 5–10 bases away from the ends. Hence, $K_{<}^{\mathrm{PM-MM,S}}$ seems to overestimate the positional effect of mismatches and thus the real situation merges aspects of both limiting cases considered. Finally, the surface attached 3′ and the free 5′ ends of the probe are not equivalent with respect to zipping and mismatches [20]. This difference is however not considered in our model.

Note that the estimation of the effect of mismatches is not only important for the problem of non-specific background subtraction but also for the quantitative analysis of target RNA with point mutations in the presence of wild-type RNA [66] and for interspecies cross-hybridization studies, for example for using human genome microarrays for non-human transcriptome analysis [67].

8. Resume: empirical description of microarray hybridization

We separately analysed several phenomena which affect the hybridization on microarrays. The system however seems too complex for aggregating all aspects discussed into one feasible model despite the approximations used. On the other hand, the results presented, in combination with recent findings, reveal some systematic trends which help to describe microarray binding data in a relatively simple, empirical fashion as a function of the probe

sequence, target concentration, the presence of mismatches and potentially also other factors such as the surface density and length of the probes, the hybridization temperature and the use of different labelling or RNA fragmentation protocols.

8.1. Effective affinities

The effective affinities of hybridization on microarrays and the corresponding absolute value of the apparent free energy are considerably smaller than those for bulk hybridization owing to surface effects such as electrostatic and entropic blocking but also because of the very broad spectrum of different sequences and lengths of probe and target oligomers present in the chip experiment which compete with each other for complex formation (see also reference [45]). Note that each of these effects multiplicatively reduces the affinity constant (see equations (2.11), (4.7), (5.2) and (5.5)), i.e.,

$$\begin{aligned} &K_{\mathrm{eff}}^{\mathrm{P-h}} \equiv K_{\mathrm{eff}}^{\mathrm{h}} \approx K^{\mathrm{P-h}} \cdot K_{\mathrm{tot}} \\ &\text{with } K_{\mathrm{tot}} = K_{\mathrm{frag}} \cdot K_{\mathrm{trunc}} \cdot K_{\mathrm{comp}} \cdot K_{\mathrm{B}} < 1. \end{aligned} \tag{8.1}$$

The affinity constants transform into the scale of free energy according to

$$G_{\mathrm{eff}}^{\mathrm{P-h}}/RT = G^{\mathrm{P-h}}/RT + \ln K_{\mathrm{tot}}. \tag{8.2}$$

Consequently the effective free energy of duplex formation is reduced by the term $\ln K_{\mathrm{tot}} < 0$ compared with the corresponding intrinsic free energy value, $G^{\mathrm{P-h}}$.

The interaction free energies of oligomers in solution were extensively investigated using well selected sequences to minimize intramolecular folding and other competing processes [14–18, 68–71]. These experiments consequently provide estimates of $G^{\mathrm{P-S}}$. The application of these solution data to hybridization on microarrays leads to the necessity to scale the free energy $G^{\mathrm{P-S}}$ down by an apparent 'temperature' of $T_{\mathrm{app}} \approx$ (600–2400) K [31, 36]. Equation (8.2) provides for this case $G^{\mathrm{P-h}}/RT_{\mathrm{app}} = G^{\mathrm{P-h}}/RT + \ln K_{\mathrm{tot}}$ and, finally, $G_{\mathrm{eff}}^{\mathrm{P-h}}/G^{\mathrm{P-h}} = T/T_{\mathrm{app}}$. Hence, the twofold–eightfold higher apparent hybridization temperature (compared with the real temperature of $T \approx 320$ K), in turn, reflects the twofold to eightfold reduction of the corresponding ideal, solution value of the free energy of duplex formation due to surface effects and competitive reactions.

A similar result was obtained by the direct comparison of the base-specific nearest neighbour free energy terms for DNA/RNA duplex formation in solution and on microarrays [40]. This study also shows that both data sets strongly correlate with each other. Surface hybridization is obviously thoroughly compatible with hybridization in solution with respect to the relative stability of the base pairings despite the marked difference of the absolute values in agreement with the theoretical results presented above.

8.2. Position dependent interaction profile

The effective strength of a base pairing in a particular molecular complex depends on its 'intrinsic' strength, i.e. the enthalpy of the bond, and on the probability that the pairing is formed. The mean interaction strength in an ensemble of probes is consequently expected to decrease towards both ends of the probe owing to zipping and RNA fragmentation or towards the free end because of the truncation of the probe. Also intramolecular folding of probe and target gives rise to a similar gradient of base pairings (unpublished results). The position dependence of the interaction profile was explicitly taken into account either by a combination of position independent, base-specific free energy contributions and a position dependent, base independent weighting function [32] or by position dependent base-specific

free energy terms [29, 41, 42, 72–74] in accordance with interaction models for hybrid duplexes in solution [15].

In the simplest case one can use a position dependent single-base (SB) description of the form

$$\log K_p^{\mathrm{P,h}}(\xi_p^{\mathrm{P}}) = -\sum_{k=1}^{L^{\mathrm{P}}} \varepsilon_k^{\mathrm{W}}(\xi_{p,k}^{\mathrm{P}}) \qquad \text{with } \varepsilon_k^{\mathrm{W}}(\xi_{p,k}^{\mathrm{P}}) = \varepsilon_{0,k}^{\mathrm{W}} + \Delta\varepsilon_k^{\mathrm{W}}(\xi_{p,k}^{\mathrm{P}}). \tag{8.3}$$

The binding constants relate to the probe type (P = PM, MM) and to specific and non-specific hybridization (h = S, NS) whereas the effective SB free energy terms, $\varepsilon_k^{\mathrm{W}}(\mathrm{B})$, estimate the stability of a complementary WC pairing (W = WC) or of a self-complementary mismatched pairing (W = SC) of base B = A, T, G or C at position k of the probe sequence, $\mathrm{B} = \xi_k^{\mathrm{P}}$. According to the discussion in the previous section one expects, for example, for the WC pairings $\varepsilon_k^{\mathrm{WC}}(\mathrm{B}) \propto g^{\mathrm{WC}}(\mathrm{B})$ and $|\varepsilon_k^{\mathrm{SC}}(\mathrm{B})| < |g^{\mathrm{WC}}(\mathrm{B})|$ (see equation (6.3)).

We applied the model to 25-meric PM and MM probes of GeneChip microarrays (see the accompanying paper [43]). The GeneChip design uses the complement of the matched base in the middle position as the mismatched base of the MM probes. It consequently forms a SC pair in the specific duplexes. The SB model (equation (8.3)) was used for a qualitative discussion of the sequence-specific effects on the observed affinities. For quantitative data analysis we however extend the description and make use of nearest neighbour or even triple-base related models to account for the stacking interactions with adjacent bases along the sequence (see [43] and [40, 72]).

8.3. Hybridization isotherm

The hybridization isotherms of microarray probes are expected to deviate from a *Langmuir*-type behaviour either in a sequence-specific fashion owing to the heterogeneous adsorption to truncated probes or in a sequence independent fashion because of electrostatic and entropic blocking (see above). In both cases the sigmoidal range of the corresponding isotherms covers a wider concentration region than a *Langmuir* isotherm with the same position of the inflection point at half-coverage, $[\mathrm{S}]_{50\%}$. The decreased slope of the isotherm with respect to the logarithmic concentration scale can be taken into account in equation (2.9) by the replacement of the linear concentration dependence by a power law, i.e.,

$$\begin{aligned} &X_p^{\mathrm{P}} = (X_p^{\mathrm{P,S}})^{a_p^{\mathrm{P}}} + X_p^{\mathrm{P,NS}} \\ &\text{with } X_p^{\mathrm{P,S}} = K_p^{\mathrm{P,S}} \cdot [\mathrm{S}] \quad \text{and} \quad X_p^{\mathrm{P,NS}} = K_p^{\mathrm{P,NS}} \cdot [\mathrm{NS}], \end{aligned} \tag{8.4}$$

where the exponent a_p^{P} is a probe-specific parameter.

Equation (2.9) with (8.4) and $X = X_p^{\mathrm{P}}$ defines the *Sips* isotherm which describes the heterogeneous adsorption to an ensemble of binding sites with a Gaussian distribution of binding free energies [75]. The *Sips* isotherm reduces to the *Langmuir* isotherm ($a_p^{\mathrm{P}} = 1$) in the limit of homogeneous adsorption where the binding free energy has a single value. Otherwise the exponent progressively decreases ($a_p^{\mathrm{P}} < 1$) as the degree of heterogeneity of the adsorption sites (i.e. the width of the distribution of binding free energies) increases. The *Sips* isotherm provides also good fits if the underlying sorption is homogeneous but concentration dependent because of electrostatic and entropic blocking (see figure 6, part (a)) and/or bulk dimerization (see equation (2.10)) and for special cases of heterogeneous adsorption where the free energies of the binding sites are however not distributed in a Gaussian fashion due to, e.g., truncated probes (equation (5.3)).

In equation (8.4) the assumption of heterogeneous binding applies only to the specific concentration because [NS] is a constant for the spiked-in data used (see [43]). In general, also, the non-specific concentration should be described by an exponential concentration

dependence. Isotherms of the *Sips* type have been proved to provide a good description of microarray hybridization [20, 37].

9. Signal intensity

The amount of probe bound RNA is detected by means of optical labels (fluorescent markers or quantum dots), which are linked, for example, to the uracyls (u*) and cytosines (c*) of the probe bound RNA fragments depending on the used labelling protocol [46]. The microarray experiment therefore measures the degree of surface hybridization of a given probe spot (indexed by p) in a scaled fashion according to (see also equation (8.4))

$$I_p^{\mathrm{P}} \approx \frac{I_{p,\max}^{\mathrm{P,S}} \cdot (X_p^{\mathrm{P,S}})^{a_p^{\mathrm{P}}} + I_{p,\max}^{\mathrm{P,NS}} \cdot X_p^{\mathrm{P,NS}}}{1 + X_p^{\mathrm{P}}} \tag{9.1}$$

if one neglects the optical background which is detected in the absence of hybridization. The degree of surface hybridization depends on the binding 'strength', $X_p^{\mathrm{P,h}}$, of the particular (perfectly matched or mismatched; P = PM, MM) DNA probe for specific (h = S) and non-specific (NS) hybridization. The scaling factors, $I_{p,\max}^{\mathrm{P,h}}$, consider the maximum fluorescence 'strength' of the hybridized RNA. The relevance of *Sips*-type and *Langmuir*-type (with $a_p^{\mathrm{P}} = 1$) equations for microarray data was previously shown in [20, 22, 30–32, 36–38, 76, 77].

The maximum fluorescence strength, $I_{p,\max}^{\mathrm{P,h}}$, is defined as the expected maximum intensity at complete surface coverage with specific (h = S) or non-specific (h = NS) RNA fragments. The fluorescence detected is related to properties of the chip such as the density of grafted oligomers and the probe area, to the amount of labelling and to the properties of the light detecting machinery (scanner, imaging software) according to $I_{p,\max}^{\mathrm{P,h}} \propto C_{\mathrm{chip}} \cdot C_{\mathrm{scan}} \cdot \langle F_p^{\mathrm{P,h}} \rangle_{\mathrm{spot}}$ (see also [39]). This simple relation neglects non-linear effects due to saturation of the detector. The labelling factor, $\langle F_p^{\mathrm{P,h}} \rangle$, in general depends on the mean number of fluorescent labels attached to the probe bound RNA fragments of the probe spot considered. The degree of labelling is a function of the RNA sequence and thus it is specific for the probe sequence and for specific and non-specific hybridization as well.

Let us assume a common labelling of PM and MM probes, $\langle F_p^{\mathrm{S}} \rangle = \langle F_p^{\mathrm{PM,S}} \rangle \approx \langle F_p^{\mathrm{MM,S}} \rangle$, and direct proportionality between the labelling of non-specific and specific fragments, $\langle F_p^{\mathrm{P,NS}} \rangle \approx r_p^F \langle F_p^{\mathrm{S}} \rangle$. The former relation is plausible because the 'specific' RNA sequence relates to the PM and to the MM probe as well giving rise to the equal number of potentially labelled bases. Differences in the number of labels of the bound fragments can however occur due to an affinity penalty of labelled bases (see below).

The second relation seems plausible because non-specific duplexes are stabilized via $L_p^{\mathrm{P,NS}} < L^{\mathrm{P}}$ WC pairings of the particular probe. If all sequence positions contribute equally to $L_p^{\mathrm{P,NS}}$ one obtains $r_p^F \approx (L^{\mathrm{RNA}} - L_p^{\mathrm{P,NS}})/(L^{\mathrm{RNA}} - L_p^{\mathrm{P}}) \approx 1$ for $L^{\mathrm{RNA}} \gg L_p^{\mathrm{P}}$, i.e., for RNA fragments which are markedly longer than the probes. Note that the mismatched, self-complementary base pairing in the specific duplexes of the MM, $\underline{\mathrm{B}}$–$\underline{\mathrm{b}}$, is replaced by a WC pair, B–b^c, in the non-specific duplexes. This substitution gives rise to an increased or decreased number of potentially labelled bases, $L_p^{\mathrm{MM,S}} = L_p^{\mathrm{MM,NS}} + 1$ for the replacement $\underline{\mathrm{B}}$–$\underline{\mathrm{b}} \rightarrow$ B–b^{c*} and $L_p^{\mathrm{MM,S}} = L_p^{\mathrm{MM,NS}} - 1$ for $\underline{\mathrm{B}}$–$\underline{\mathrm{b}}^* \rightarrow$ B–b^c, respectively. This difference of one potentially labelled base can however be ignored in the limit of long RNA fragments.

On the basis of these arguments we assume a common fluorescence strength of specific and non-specific RNA fragments which transforms equation (9.1) into (see also equation (8.4))

$$I_p^{\mathrm{P}} \approx I_p^{\max} \cdot \theta_p^{\mathrm{P}} \qquad \text{with } \theta_p^{\mathrm{P}} = \frac{X_p^{\mathrm{P}}}{1 + X_p^{\mathrm{P}}}. \tag{9.2}$$

Accordingly, the binding isotherm of the probe considered is simply scaled by a probe-specific intensity factor.

The variability of the maximum signal intensity among all probe spots due to the labelling effect can be estimated assuming a random base composition of the RNA fragments with a fraction of potentially labelled bases of $x_{\mathrm{lab}} \approx 0.5$ (e.g., if labels are attached only to the cytosines and uracyls) which are distributed according to a binomial distribution. The coefficient of variation of the potentially labelled bases, $\mathrm{CV}(L^{\mathrm{lab}}) \approx \{(1 - x_{\mathrm{lab}})/(x_{\mathrm{lab}} \cdot L^{\mathrm{RNA}})\}^{0.5} \approx (L^{\mathrm{RNA}})^{-0.5}$, provides a measure of the relative variation of the number of labelled bases between different probes and thus also of the variability of $I_p^{\max}$. One obtains $\mathrm{CV}(L^{\mathrm{lab}}) \approx 0.15$–$0.05$ for RNA fragments of length $L^{\mathrm{RNA}} = 50$–500. In other words, the variation of the labelling among the different probe sequences are expected to affect the maximum intensity value on the average only weakly by, at maximum, a few per cent of its mean value.

We so far assumed that the hybridization yield is independent of the presence of labels. Previous studies suggest, however, that the attached labels hamper duplex formation [29, 40]. Note that the individual RNA fragments of one sequence and length can carry different numbers of labels with a certain distribution about its average value relating to the case where labelling does not affect the binding affinity, $F_p^{\mathrm{P},0} \propto p_{\mathrm{lab}} \cdot L_p^{\mathrm{P}}$ (where $p_{\mathrm{lab}} \approx 1/10 < 1$ is the labelling yield, i.e., the probability of attaching a fluorescent label, for example, to a cytosine or uracyl in the RNA sequence [46]). The hybridization reaction selects the RNA fragments with a relatively small number of labels to minimize the affinity penalty of the target region upon duplex formation [29]. As a consequence, the average number of labels per bound transcript is smaller than the random mean, $\langle F_p^{\mathrm{P}} \rangle < F_p^{\mathrm{P},0}$, and it decreases with increasing affinity penalty per label. Only RNA fragments without labels inside the target region of length L^{P} are expected to bind to probes in significant amounts in the limit of a high penalty.

In summary, the intensity of a probe spot is governed by two sequence-specific factors: the sequence of the target region of length L^{P} mainly determines the affinity for DNA/RNA duplex formation and thus the binding strength of a given probe whereas the RNA sequence outside of the target region of length $L^{\mathrm{RNA}} - L^{\mathrm{P}}$ mainly determines the average number of labels per duplex and thus its fluorescence strength. The former affinity factor is expected to dominate the intensity difference between different probes because it depends exponentially on the free energy which in turn is linearly related to its base composition. In contrast, the fluorescence factor is a relatively weak, linear function of the sequence with a relatively small variability between the probes.

10. Kinetic effects and non-equilibrium thermodynamics

The hybridization effects discussed relate to thermodynamic equilibrium if sorption and desorption rates are equal in magnitude. Usually the hybridization experiment starts by adding the solution of RNA transcripts onto the 'empty' chip with free probes. In the first stage sorption prevails over desorption until both processes reach the steady state. Equilibration typically lasts dozens of hours to allow the system to reach the steady state, and is followed by fluorescent labelling and subsequent washing to remove unbound RNA before scanning.

The binding kinetics in a complex sample was found to take significantly longer for specific than for non-specific binding [76]. It has been suggested that one use the different kinetics to estimate and to correct for the hybridization contributed by non-specific binding. Model calculations of the effects of diffusion, cross-hybridization, relaxation time and target concentration on the hybridization kinetics show that the presence of non-specific RNA

profoundly slows down the equilibration time of specific targets on a timescale comparable to the time of the experiment and in this way potentially confounds interpretation of the data [26]. Another recent theoretical study argues that the ratio of non-specific and specific hybrids can change dramatically with time and can differ considerably from the equilibrium concentrations of bound species [78].

Peterson suggests that the 'apparent' asymptotic intensity value reached might be an artefact due to the limited time window of the experiment [19]. On the other hand, non-equilibrium models have been found to improve the fit of GeneChip microarray data only marginally compared with the equilibrium *Langmuir*-type model [37]. This result implies that the *Langmuir* and/or *Sips* isotherms formally apply also to hybridizations which did not reach equilibrium. In this case, the effective binding constants extracted deviate however from their equilibrium values and thus their interpretation is complicated by an additional factor.

11. Summary and conclusions

We studied thermodynamic aspects of surface adsorption of RNA transcripts to DNA oligonucleotide probes grafted on microarrays. The theoretical analysis establishes the relation between the amount of bound RNA and the concentration of specific target RNA in the supernatant solution for different situations which consider effects such as non-specific hybridization, bulk dimerization and other competitive processes, the electrostatic and entropic repulsion between surface tethered probes and dissolved transcripts, the truncation of probe and target and the zipping of probe/target duplexes. The factors considered affect the concentration of half-coverage, the residual coverage in the absence of specific transcripts and the shape of the sorption isotherm mostly in a sequence-specific fashion. Isotherms of the *Langmuir* or *Sips* type are predicted to provide a relatively simple description of the non-linear, probe-specific concentration dependence of the signal intensity of microarray probes.

The microarray experiment intends to measure the degree of expression of the target gene in terms of the concentration of the specific transcript. The signal analysis consequently requires the correction of the measured intensity for the probe-specific affinity of specific hybridization, for non-linear effects due to saturation at higher concentrations and for the 'chemical background' owing to non-specific hybridization, to obtain reliable estimates of the RNA concentration. In the accompanying paper we address this issue using experimental microarray intensity data and the sequences of the corresponding microarray probes (see [43]).

Acknowledgment

The work was supported by the Deutsche Forschungsgemeinschaft under grant no BIZ 6-1/3.

References

[1] Schena M 1996 *Bioessays* **18** 427
[2] Lipshutz R J, Fodor S P A, Gingeras T R and Lockhart D J 1999 *Nat. Genet.* **21** 20
[3] Holstege F C P, Jennings E G, Wyrick J J, Lee T I, Hengartner C J, Green M R, Golub T R, Lander E S and Young R A 1998 *Cell* **95** 717
[4] Lee C-K, Klopp R G, Weindruch R and Prolla T A 1999 *Science* **285** 1390
[5] Golub T R *et al* 1999 *Science* **286** 531
[6] Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D and Levine A J 1999 *Proc. Natl Acad. Sci. USA* **96** 6745
[7] Hippo Y, Taniguchi H, Tsutsumi S, Machida N, Chong J-M, Fukayama M, Kodama T and Aburatani H 2002 *Cancer Res.* **62** 233

[8] Clarke P A, te Poele R, Wooster R and Workman P 2001 *Biochem. Pharmacol.* **62** 1311
[9] Debouck C and Goodfellow P N 1999 *Nat. Med.* **21** 48
[10] Cheng J *et al* 2005 *Science* **308** 1149
[11] Kapranov P, Cawley S E, Drenkow J, Bekiranov S, Strausberg R L, Fodor S P A and Gingeras T R 2002 *Science* **296** 916
[12] Hong B-J, Oh S-J, Youn T-O, Kwon S-H and Park J-W 2005 *Langmuir* **21** 4257
[13] Affymetrix 2001 *User Guide* (Santa Clara, CA: Affymetrix, Inc.)
[14] Bloomfield V A, Crothers D M and Tinoco I 2000 *Nucleic Acids Structures, Properties and Functions* (Sausalito: University Science Books)
[15] Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M and Sasaki M 1995 *Biochemistry* **34** 11211
[16] Sugimoto N, Nakano S, Yoneyama M and Honda K 1996 *Nucleic Acids Res.* **24** 4501
[17] Sugimoto N, Nakano M and Nakano S 2000 *Biochemistry* **39** 11270
[18] Wu P, Nakano S and Sugimoto N 2002 *Eur. J. Biochem.* **269** 2821
[19] Peterson A W, Heaton R J and Georgiadis R M 2001 *Nucleic Acids Res.* **29** 5163
[20] Peterson A W, Wolf L K and Georgiadis R M 2002 *J. Am. Chem. Soc.* **124** 14601
[21] Watterson J H, Piunno P A E, Wust C C and Krull U J 2000 *Langmuir* **16** 4984
[22] Halperin A, Buhot A and Zhulina E B 2004 *Biophys. J.* **86** 718
[23] Naef F, Lim D A, Patil N and Magnasco M 2002 *Phys. Rev.* E **65** 4092
[24] Chou C-C, Chen C-H, Lee T-T and Peck K 2004 *Nucleic Acids Res.* **32** e99
[25] Chudin E, Walker R, Kosaka A, Wu S, Rabert D, Chang T and Kreder D 2001 *Genome Biol.* **3** research 0005.1
[26] Bhanot G, Louzoun Y, Zhu J and DeLisi C 2003 *Biophys. J.* **84** 124
[27] Chan V, Graves D and McKenzie S 1995 *Biophys. J.* **69** 2243
[28] Steel A B, Levicky R L, Herne T M and Tarlov M J 2000 *Biophys. J.* **79** 975
[29] Naef F and Magnasco M O 2003 *Phys. Rev.* E **68** 11906
[30] Hekstra D, Taussig A R, Magnasco M and Naef F 2003 *Nucleic Acids Res.* **31** 1962
[31] Held G A, Grinstein G and Tu Y 2003 *Proc. Natl Acad. Sci. USA* **100** 7575
[32] Zhang L, Miles M F and Aldape K D 2003 *Nat. Biotechnol.* **21** 818
[33] Dimitrov R A and Zuker M 2004 *Biophys. J.* **87** 215
[34] Vainrub A and Pettitt B M 2002 *Phys. Rev.* E **66** 041905
[35] Matveeva O V, Shabalina S A, Nemtsov V A, Tsodikov A D, Gesteland R F and Atkins J F 2003 *Nucleic Acids Res.* **31** 4211
[36] Carlon E and Heim T 2004 *Preprint* q-bio.BM/0411011 v1
[37] Burden C J, Pittelkow Y E and Wilson S R 2004 *Stat. Appl. Genet. Mol. Biol.* **3** 35
[38] Burden C J, Pittelkow Y E and Wilson S R 2004 *Preprint* q-bio.BM/0411005 v1
[39] Binder H, Kirsten T, Loeffler M and Stadler P 2004 *J. Phys. Chem.* B **108** 18003
[40] Binder H, Kirsten T, Hofacker I, Stadler P and Loeffler M 2004 *J. Phys. Chem.* B **108** 18015
[41] Binder H, Preibisch S and Kirsten T 2005 *Langmuir* **21** 9287
[42] Mei R *et al* 2003 *Proc. Natl Acad. Sci. USA* **100** 11237
[43] Binder H and Preibisch S 2006 *J. Phys.: Condens. Matter* **18** S537
[44] Mathews D H, Burkard M E, Freier S M, Wyatt J R and Turner D H 1999 *RNA* **5** 1458
[45] Levicky R and Horgan A 2005 *Trends Biotechnol.* **23** 143
[46] Affymetrix 2004 IVT Labeling Kit, *Technical Note* 1
[47] Cantor C R and Schimmel P R 2002 *Biophysical Chemistry* (New York: Freeman)
[48] Barone F, Cellai L, Matzeu F and Pedone F 2000 *Biophys. Chem.* **86** 37
[49] Deutsch J M, Liang S and Narayan O 2004 *Preprint* q-bio.BM/0406039 v1
[50] Relogio A, Schwager C, Richter A, Ansorge W and Valcarcel J 2002 *Nucleic Acids Res.* **30** e51
[51] Tolstrup N, Nielsen P S, Kolberg J G, Frankel A M, Vissing H and Kauppinen S 2003 *Nucleic Acids Res.* **31** 3758
[52] Luebke K J, Balog R P and Garner H R 2003 *Nucleic Acids Res.* **31** 750
[53] Binder H and Lindblom G 2003 *Phys. Chem. Chem. Phys.* **5** 5108
[54] Halperin A, Buhot A and Zhulina E B 2005 *Biophys. J.* **89** 796
[55] Hughes T R *et al* 2001 *Nat. Biotechnol.* **19** 342
[56] Granjeaud S, Bertucci F and Jordan B R 1999 *Bioessays* **21** 781
[57] Jeon Y, Bekiranov S, Karnani N, Kapranov P, Ghosh S, MacAlpine D, Lee C, Hwang D S, Gingeras T R and Dutta A 2005 *Proc. Natl Acad. Sci. USA* **102** 6419
[58] Cawley S, Bekiranov S, Ng H H, Kapranov P, Sekinger E A, Kampa D, Piccolboni A, Sementchenko V, Cheng J and Williams A J 2004 *Cell* **116** 499
[59] McGall G H, Barone A D, Diggelman M, Fodor S P A, Gentalen E and Ngo N 1997 *J. Am. Chem. Soc.* **119** 5081

[60] McGall G, Labadie J, Brock P, Wallraff G, Nguyen T and Hinsberg W 1996 *Proc. Natl Acad. Sci. USA* **93** 13555
[61] Pirrung M C and Fallon L 1998 *J. Org. Chem.* **63** 241
[62] Jobs M, Fredriksson S, Brookes A J and Ulf L 2002 *Anal. Chem.* **74** 199
[63] Blossey R and Carlon E 2003 *Phys. Rev.* E **68** 061911
[64] Dorris D R, Nguyen A, Gieser L, Lockner R, Lublinsky A, Patterson M, Touma E, Sendera T J, Elghanian R and Mazumder A 2003 *BMC Biotechnol.* **3** 6
[65] Lee I, Dombkowski A A and Athey B D 2004 *Nucleic Acids Res.* **32** 681
[66] Halperin A, Buhot A and Zhulina E B 2004 *Clin. Chem.* **50** 2254
[67] Enard W *et al* 2002 *Science* **296** 340
[68] Wu P and Sugimoto N 2000 *Nucleic Acids Res.* **28** 4762
[69] Peyret N, Seneviratne P A, Allawi H T and SantaLucia J 1999 *Biochemistry* **38** 3468
[70] Xia T, SantaLucia J, Burkard M E, Kierzek R, Schroeder S J, Jiao X, Cox C and Turner D H 1998 *Biochemistry* **37** 14719
[71] Bommarito S, Peyret N and SantaLucia J 2000 *Nucleic Acids Res.* **28** 1929
[72] Binder H, Kirsten T, Loeffler M and Stadler P 2003 *Proc. German Bioinformatics Conf.* vol 2, p 145
[73] Binder H 2006 *Bioinformatics of Gene Regulation* vol II, ed N Kolchanov, R Hofestaedt and L Milanesi, Springer Sciences and Business Media, p 451
[74] Binder H and Preibisch S 2005 *Biophys. J.* **89** 337
[75] Sips R 1948 *J. Phys. Chem.* **16** 490
[76] Dai H, Meyer M, Stepaniants S, Ziman M and Stoughton R 2002 *Nucleic Acids Res.* **30** e86
[77] Kepler T B, Crosby L and Morgan K T 2002 *Genome Biol.* **3** RESEARCH0037
[78] Zhang Y, Hammer D A and Graves D J 2005 *Biophys. J.* **89** 2950