JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 5, 2004

Pp. 945–970

Towards Quality Control for DNA Microarrays

KALEIGH SMITH and MIKE HALLETT

ABSTRACT

We present a framework for detecting probes in oligonucleotide microarrays that may add significant error to measurements in hybridization experiments. Four types of so-called *degenerate* probe behavior are considered: secondary structure formation, self-dimerization, cross-hybridization, and dimerization. The framework uses a well-established model for computing the free energy of nucleic acid sequence hybridization and a novel method for the detection of patterns in hybridization experiment data. Our primary result is the identification of unique patterns in hybridization experiment data that are shown to correlate with each type of degenerate probe behavior. A support function for identifying degenerate probes from a large set of hybridization experiments is given and some preliminary experimental results are given for the Affymetrix HuGeneFL GeneChip. Finally, we show a strong relationship between the Affymetrix *discrimination* measure for a probe and the free-energy estimate from theoretical models of hybridization. In particular, probes on the HuGeneFL GeneChip with high free-energy estimates (weak hybridization) have almost always approximately zero discrimination. The framework can be applied to any Affymetrix oligonucleotide array, and the software is made freely available to the community.

Key words: microarrays, cross-hybridization, secondary structure, dimerization.

1. INTRODUCTION

TECHNOLOGIES SUCH AS AFFYMETRIX OLIGONUCLEOTIDE ARRAYS (DNA chips or GeneChips) have launched several new fields in functional genomics. However, the amount of error in the measurements of gene expression produced by these technologies represents a significant obstacle to understanding gene regulatory mechanisms. This paper presents a framework for identifying several types of error common to Affymetrix oligonucleotide arrays.

Essentially, an organism specific array (a.k.a. *chip*) contains a set of *probes* that target most, if not all, genes and other areas of biological interest (termed *targets*) in the genome of the organism. Each probe p is an oligonucleotide of length 20–25 bp and is tethered to the chip at a specific location. The nucleic acid sequence of each such probe is the Watson–Crick complement of a nucleotide sequence t that is located ideally in exactly one position of one target in the genome of the organism. We term this complementary strand t a *tag* and say that tag t *matches* probe p. Each target is typically represented by between 11 and 60 such tags.

A *hybridization experiment* consists of harvesting, under some specific condition, a sufficiently large sample of mRNA transcripts from the tissue or organism under study. In an experiment, the mRNA is

McGill Centre for Bioinformatics, McGill University, Montreal, Canada.

preprocessed and labeled to form a sample of cRNA. This sample is stained and brought into contact with the chip. Those cRNA tags present in the sample should then hybridize with (and only with) their matching probes on the chip. The intensity of each RNA/DNA hybrid (termed *probe–tag pair*) is optically measured. In an error-free scenario, this intensity is proportional to the true number of transcripts present in the sample. A statistically robust method is then applied to the set of probe–tag pairs representing each target to estimate the quantitative level of expression for that target. See, for example, Affymetrix (2002), Li and Wong (2001a), and Li and Wong (2001b).

Variability among different hybridization experiments is commonly characterized as biological variability, sample variability, or technical variability. We are primarily concerned here with technical variability. Technical variability introduces error to hybridization experiments in the form of *stochastic error* that may be caused by lab equipment or conditions, or *bias error* that may be caused by the design or construction of the chip. Our concern here is with the detection (and eventually prediction) of bias error in experiments. Several strategies exist in the literature for designing oligonucleotide microarrays that attempt to minimize (or detect) technical bias error (BenDor *et al.*, 2000; Hubbell and Pevzner, 1999; Sengupta and Tompa, 2000; Tobler *et al.*, 2002). These methods, however, focus primarily on establishing design rules for the *de novo* construction of universal microarrays or on the inclusion of probes to detect faults in the chip construction phase. The work presented here introduces a framework to predict when probes and their matching tags will exhibit *degenerate behavior* (their intensity readings are consistently not proportional to the true number of transcripts present in the sample). The framework uses both probe sequence information supplied by NetAffx (2001) and a large set of hybridization experiments. The framework will provide us with a tool for the *in silico* evaluation of the quality of a chip before it is manufactured. The predictions also act as a filter to remove some types of error common to gene expression studies.

This paper considers four types of degenerate behavior that we conjecture to add a significant amount of error to intensity measurements in hybridization experiments: secondary structure formation (a tag or probe strongly hybridizes with itself), self-dimerization (two copies of the same tag hybridize), dimerization (two distinct tags in the sample hybridize), and cross-hybridization (two distinct tags t, t' with matching probes p, p', respectively, tend to hybridize with both p and p'). A well-designed microarray should prevent the occurrence of these four types of degeneracy.

Each of the four degenerate behaviors may contribute to error in gene expression measurements in a distinct way. As a simple example, consider a set of tags $t_1, \ldots, t_l$ representing some target g where t_1 is known to form a strong secondary structure. We assume further that the remaining tags do not have an affinity to form secondary structure. Let $p_1, \ldots, p_l$ be the matching probes for $t_1, \ldots, t_l$, respectively. During a hybridization experiment, t_1 will have a tendency to hybridize with itself, and therefore it will tend not to hybridize with its matching probe p_1 on the chip at the same rate as t_i, $i > 1$. The intensities of probe–tag pair (p_1, t_1) would then be consistently lower than would be witnessed had a better tag been chosen. Moreover, the intensities for this probe–tag pair should tend to be consistently lower than other probe–tag pairs representing the same target. Therefore, the intensity of the entire target (computed as a weighted average of the intensities of all tags representing that target) will tend to be lower than the true number of transcripts present in the sample. We explore similar style arguments for the remaining types of degeneracy in this paper.

For each of the four types of degeneracy, we conjecture that the pattern of intensity measurements for degenerate probes over a sufficiently large set of hybridization experiments is distinct from the pattern of intensity measurements for nondegenerate probes. This distinct pattern for degenerate probes is recognizable by comparing the intensity of a probe to the intensities of the remaining probes in its probe group. For example, a plausible pattern of degenerate behavior caused by secondary structure is illustrated as follows. Consider once again a set of tags $t_1, \ldots, t_l$ representing some target g. If the intensity measurements of the probe p_1 associated with t_1 rank extremely low w.r.t. the intensity measurements for probes associated with $t_2, \ldots, t_l$, especially when the target g is highly expressed, then it is possible that t_1 is prone to form secondary structure and therefore is not hybridizing with p_1 at the same rate as the other tags hybridize with their respective probes. When the target g is lowly expressed, we could expect that the rank of the intensity for probe p_1 would be more uniformly distributed between $1..l$.

The above example begs the following questions: (i) *How can we predict whether a tag or probe has an affinity to exhibit degenerate behavior*? (ii) *What is the appropriate definition for the pattern of intensity measurements for nondegenerate and degenerate probes*? (iii) *How do we detect these patterns and measure the significance of such putative degenerate patterns*?

We begin in Section 2 by defining a *conflict graph* for a chip. Each vertex in the conflict graph corresponds to a probe on the chip. Edges in the conflict graph correspond to degenerate behavior of a probe (or the corresponding tag) or pair of probes (the corresponding pair of tags). With respect to question (i), the affinity for a probe (or pair of probes) to exhibit each type of degeneracy is estimated via a well-studied method based on theoretical models for calculating the free-energy (ΔG) of a hybridization for either a single sequence (secondary structure) or between two sequences (cross-hybridization, dimerization) (Mathews *et al.*, 1999; Sankoff and Zuker, 1984; SantaLucia, 1998a, 1998b, 2002; Ship *et al.*, 2002; Zuker *et al.*, 1999). An edge is added to the conflict graph if and only if the minimum free energy of hybridization that causes degeneracy is below a conservative threshold.

Section 3 uses the Affymetrix HuGeneFL GeneChip (NetAffx 2002) and a set of 126 hybridization experiments (i.e., cell files) from three separate laboratories. The HuGeneFL GeneChip is a relatively old chip for which there are many hybridization experiments available to the general public. We examine the structure of the conflict graph induced by the HuGeneFL chip and our theoretical model of hybridization. We find several interesting facts including that the putative degenerate probes tend to have low ΔG estimates, the conflict graph tends have many low degree vertices, and furthermore, that the conflict graph is highly disconnected.

W.r.t. question (ii), Section 4 gives both an intuitive justification and a formal definition of the patterns for degenerate and nondegenerate probe behavior in hybridization experiments. We begin by showing that the set of predicted nondegenerate probes displays a pattern of intensity measurements distinct from the pattern of intensity measurements for probes predicted to be degenerate. We also show that the set of probes predicted via the theoretical hybridization model to exhibit degenerate behavior do in fact follow our pattern for degenerate probe behavior. More precisely, the distribution of ranks for a probe predicted to have an affinity for a specific degenerate behavior is different than the distribution of ranks for nondegenerate probes (the background distribution), and furthermore, this distribution of ranks agrees with our intuition.

As a chip is increasingly used by the community and the raw intensity values (i.e., the cell files) are made publicly available, this library of knowledge should give us the ability to distinguish between degenerate and nondegenerate probes on the chip. To realize such a strategy, we investigate how well the patterns of degenerate and nondegenerate probes (for each type of degenerate behavior) can be used alone to measure the affinity for a probe to exhibit degenerate behavior. In Section 5, we define support functions for estimating the log-likelihood ratio that a probe has an affinity for a specific type of degenerate behavior (secondary structure, dimerization, cross-hybridization) given only a large set of hybridization experiments. The results in Section 5 indicate that many additional hybridizations are required, if our support functions are to function correctly. We provide a weak lower bound on this number in Section 5.2.

In Section 6, we incorporate *discrimination* into our framework and investigate its relationships with our estimates of free energy for hybridizations. The notion of *discrimination* introduced by Affymetrix measures the ability of the intensity for a probe to represent the true number of mRNA transcripts present in a sample. Essentially, the ratio of intensities for a perfect match probe and a mismatch probe is used to estimate the specificity of the probe. We give experimental evidence of a correlation between probe–tag pairs with high ΔG estimates (weak hybridization properties) and discrimination values that are almost always approximately equal to zero (the perfect match and mismatch intensities are equal).

Finally, in Section 7, we state a number of open problems and future directions.

2. NOTATION AND TOOLS

2.1. Probes, tags, and groups

Let $\Sigma_{dna} = \{A, C, G, T\}$ and $\Sigma_{rna} = \{A, C, G, U\}$. A *probe* p of length n is a string $p = p_1 p_2 \dots p_n \in \Sigma_{dna}^n$. A *tag* t of length n is a string $t \in \Sigma_{rna}^n$. The *reverse* t^r of t is the string $t_n t_{n-1} \dots t_1$. The *Watson–Crick (wc-) complement* $\bar{q}$ of q is the string obtained from q^r by interchanging $A \leftrightarrow T$ and $C \leftrightarrow G$. We say that two strings s and t are a *wc-complementary* iff $\bar{s} = t$. We say that probe p *matches* a tag t (or t matches p) iff t is the wc-complement of p after replacing U with T.

Let $T = \{t_1, \dots, t_l\}$ be a set of tags, $t_i \in \Sigma_{rna}^n$. Let $P = \{p_i : p_i \text{ matches } t_i \in T\}$ be the set of corresponding probes. The probes are fixed to the chip whilst the tags are derived from the mRNA in the sample. Let $G = \{g_1, \dots, g_m\}$ be the set of targets (genes or other areas of biological significance).

For our purposes, each $g \in G$ is represented by a (unique) set of tags $T_g \subseteq T$. For all $g, g' \in G$, $g \neq g'$, $T_g \cap T_{g'} = \emptyset$. Let P_g represent the set of probes which match the set of tags T_g; i.e., $P_g = \{p \in P : p \text{ matches some } t \in T_g\}$. We call T_g the *tag group* and P_g the *probe group* for g and T_g and P_g are said to *target* g.

Definition 1 (Chip). *A* chip $C = \langle G, P, \{P_g : g \in G\}, T, \{T_g : g \in G\}\rangle$ *is comprised of a group* G *of genes, sets* P *and* T *of probes and tags, and sets of probe groups* $P_g \in P$ *and tag groups* $T_g \in T$ *for each gene* $g \in G$.

2.2. Models of hybridization

Affinity for duplex formation is most commonly measured in terms of duplex stability or hybridization strength by *free energy* ΔG in kcal/mol (SantaLucia, 1998). Throughout this paper, all ΔG measurements are in kcal/mol. This is defined as the total change in energy from duplex to single-stranded states. In a series of papers (SantaLucia *et al.*, 1996, 1998a, 1998b, 2002), the thermodynamic hybridization parameters for most DNA 2-mers against most 2-mers including both wc-complementary 2-mers and mismatch 2-mers, and various misalignments were determined. These parameters are used with the *nearest-neighbor* model (N-N) for calculating the ΔG for a pair of (not necessarily wc-complement) nucleic acid sequences. The N-N model predicts the ΔG of an alignment of a pair of nucleic acid sequences by summing the thermodynamic hybridization parameters for each occurring 2-mer against the 2-mer to which it is aligned and thermodynamic parameters for the energy required to initiate duplex formation. Essentially, lower ΔG scores for two nucleic acid sequences indicate a stronger hybridization between the nucleic acid sequences. The N-N model is believed to give accurate predictions of duplex free energy for nucleic acid sequences of length 5–60 (SantaLucia *et al.*, 1996).

Observation 1. *The minimum* ΔG *over all alignments between nucleic acid sequences* $t = t_1 \ldots t_n$ *and* $s = s_1 \ldots s_m$ *can be found in* $O(nm)$ *time and* $O(n + m)$ *space.*

The algorithm denoted by $\mathcal{DP}$ uses standard dynamic programming with a constant gap penalty. As a bulge between two short nucleic acid sequences is unlikely, the internal-gap penalty is extremely high. The algorithm takes into consideration ionic and temperature conditions. The computation of free energy between two tags $t, t' \in T$ is an RNA versus RNA alignment whereas the computation of free energy between a probe $p \in P$ and a tag $t \in T$ is a DNA versus RNA alignment. Since the complete parameters, including mismatch parameters, required to compute the free energy for RNA versus DNA and RNA versus RNA alignments were not publicly available, we use the parameters for DNA versus DNA alignments as a good approximation (Sugimoto and Nakano, 1996). Therefore, input to $\mathcal{DP}$ may be over Σ_{dna} or Σ_{rna}; however, in the latter case, the sequence is translated from Σ_{rna} to Σ_{dna} by replacing U with T. We realize that the N-N model applied to the wc-complement of probes does not take into account the "trailing sequence" of an actual cRNA tag in a sample. We also realize that DNA/DNA parameters will not give us exact measurements for DNA/RNA and RNA/RNA structures. This is acceptable, as we require a relative measure of stability, not an absolute measure.

We implemented an algorithm for the prediction of secondary structure that is built upon the standard method for the prediction of RNA secondary structure (Sankoff and Zuker, 1984) with the DNA parameter set of SantaLucia *et al.* and the N-N model. We assume that more complicated secondary structures will not form in short sequences (length 25). Under this assumption, our program returns the minimum free energy over all hairpin structures (without pseudo-knots) for a nucleic acid sequence. We represent this as function $\mathcal{S} : \Sigma_{dna}^n \to \mathbb{R}$; $\mathcal{S}(p)$ is the affinity for probe p to form secondary structure. Let t be the tag matching p. It is clearly the case that the affinity for t to form secondary structure will differ from the affinity for p to form secondary structure for several reasons: (i) the probe p is tethered to the chip while tag t is not, (ii) tag $t \in \Sigma_{rna}^*$, $p \in \Sigma_{dna}^*$, and (iii) the free energy of a secondary structure formation for t will be affected by a "tailing" ribonucleic acid sequence. However, for ease of exposition, we use $\mathcal{S}(p)$ as an approximation for the affinity for tag t to form secondary structure. That is, we assume $\mathcal{S}(p) \approx \mathcal{S}(t)$. We write that a probe p has high affinity to form secondary structure to mean either t or p may form secondary structure.

We use the algorithm $\mathcal{DP}$ to predict the pairwise behavior of two distinct probes $p, p' \in P$. Let $t, t' \in T$ be the respective matching tags of p and p'. For p, p', the function for cross-hybridization $\mathcal{X}(p, p') = min(\mathcal{DP}(p, t'), \mathcal{DP}(p', t))$ computes the affinity for t' to hybridize with p and t to hybridize with p'. Similarly, for dimerization, let $\mathcal{D}(p, p')$ be the result of computing $\mathcal{DP}(t, t')$ where t, t' matches p, p', respectively. In the case of dimerization, we allow that $p = p'$. For ease of notation, we say that probe p has an affinity to self-dimerize. A probe p is degenerate if any value for $\mathcal{S}(p)$, $\mathcal{D}(p, p)$, $\mathcal{X}(p, p')$, or $\mathcal{D}(p, p')$ indicates an affinity for one of the degeneracy causing hybridization behaviors.

2.3. Conflict graph for a chip

To organize the properties of probes and relationships between probes on a chip we use a *conflict graph*. Given a chip $C = \langle G, P, \{P_g : g \in G\}, T, \{T_g : g \in G\}\rangle$, we create an edge labeled multigraph $M = (V, E, \tau, \kappa)$. Each probe $p \in P$ on the chip corresponds to a vertex $p \in V$. Here, κ is a function labeling the edges of M, $\kappa : E \to \{s, x, d\}$, and $\tau = \{\tau_s, \tau_x, \tau_d\}$ is a set of suitable threshold parameters for the $\mathcal{S}$, $\mathcal{X}$, $\mathcal{D}$ functions from Section 2.2.

Formally, we include an edge $(p, p) \in E$, $p \in P$, if $\mathcal{S}(t) < \tau_s$, where t matches p and set $\kappa(p, p) \leftarrow s$. If $\mathcal{X}(p, p') < \tau_x$, then we include an edge $(p, p') \in E$ and set $\kappa(p, p') \leftarrow x$. If $\mathcal{D}(p, p') < \tau_d$, then we include an edge $(p, p') \in E$ and set $\kappa(p, p') \leftarrow d$. In the case of *self-dimerization* (two copies of t dimerize), we add a self-loop to p and assign $\kappa(p, p) \leftarrow d$. We may also wish to introduce a threshold for self-dimerization τ_{sd} that differs from τ_d. We describe how to chose the ΔG thresholds, τ_s, τ_x, τ_d, experimentally with respect to chip C in Section 3.1. These sets of secondary structure, self-dimerization, cross-hybridization, and dimerization edges are denoted by S, SD, X, and D, respectively.

3. ANALYSIS OF AFFYMETRIX HUGENEFL GENECHIP

Our experiments are based on a set of 126 Affymetrix Inc. (TM) hybridization experiments made publicly available by Lemon *et al.* (2002), Ship *et al.* (2002), and Virtaneva *et al.* (2001). These hybridization experiments used the Affymetrix HuGeneFL chip (TM); this is a human-specific chip and has a probe set P' of size 131,542 used to represent a set G' of 7,129 genomic targets or groups. We chose the HuGeneFL chip since it targets a significantly high number of genes, the probe sets (and sequence) are available at NetAffx (2002), and a large set of gene expression datasets are publicly available. Affymetrix defines a probe group P_g for each $g \in G'$. For our experiments, we consider only those probe groups of size 20 from HuGeneFL. That is, chip C has targets $G = \{g \in G' : |P_g| = 20\}$ and probes $P = \{p \in P' : p \in P_g \text{ for some } g \in G\}$. The tag set T, tag groups $\{T_g\}$, and probe groups $\{P_g\}$ are formed from G and P. After these restrictions and the removal of several Affymetrix control groups, we are left with $|G| = 6{,}378$ and $|P| = 127{,}560$ of which 127,386 are unique DNA sequences. (However, G does contain 58 Affymetrix control groups.)

3.1. Analysis of ΔG: Choosing thresholds

This subsection gives our rationale for choosing the thresholds τ_s, τ_d, and τ_x for deciding whether a probe has an affinity for secondary structure, dimerization, or cross-hybridization. We begin by studying the function $\mathcal{DP}(p, \bar{p})$ for the calculation of free energy (ΔG) for a randomly chosen set of 10^8 probes $p \in \Sigma_{dna}^{25}$ and their wc-complements $\bar{p}$. We do this in order to determine whether the distribution of ΔG over a large random set of length 25 probes is the same as the distribution of ΔG for the specific probes on the HuGeneFL chip. We find that $\mathcal{DP}(p, \bar{p})$ is normally distributed about a mean of -29.2981 with minimum and maximum values of -45.083 and -15.603 over all p (Fig. 1 in Smith and Hallett [2004]). The average ΔG for the set of all probe–tag pairs for the HuGeneFL chip C is -29.1549 with range of -42.003 and -18.993. From Fig. 2 in Smith and Hallett (2004), we conclude that the random set of probes is a good approximation to the probes in HuGeneFL. For chip C, let τ_c be the ΔG for the weakest probe–tag pair. For HuGeneFL, $\tau_c \approx -19$. We note that for newer chips, the calculation of the thresholds should be redone, since newer probe construction strategies bias the set of probes towards having higher GC content. We now compute $\mathcal{X}(p, p')$ and $\mathcal{D}(p, p')$ for all distinct $p, p' \in P$. In total, 239 of approximately 8×10^9 pairs of probes have $\mathcal{X}(\cdot, \cdot) \leq \tau_c$ and, in total, 487 such pairs have $\mathcal{D}(\cdot, \cdot) \leq \tau_c$. Figures 1(a) and (b) depict the distribution of $\mathcal{X}(\cdot, \cdot)$ and $\mathcal{D}(\cdot, \cdot)$ for such pairs. F1

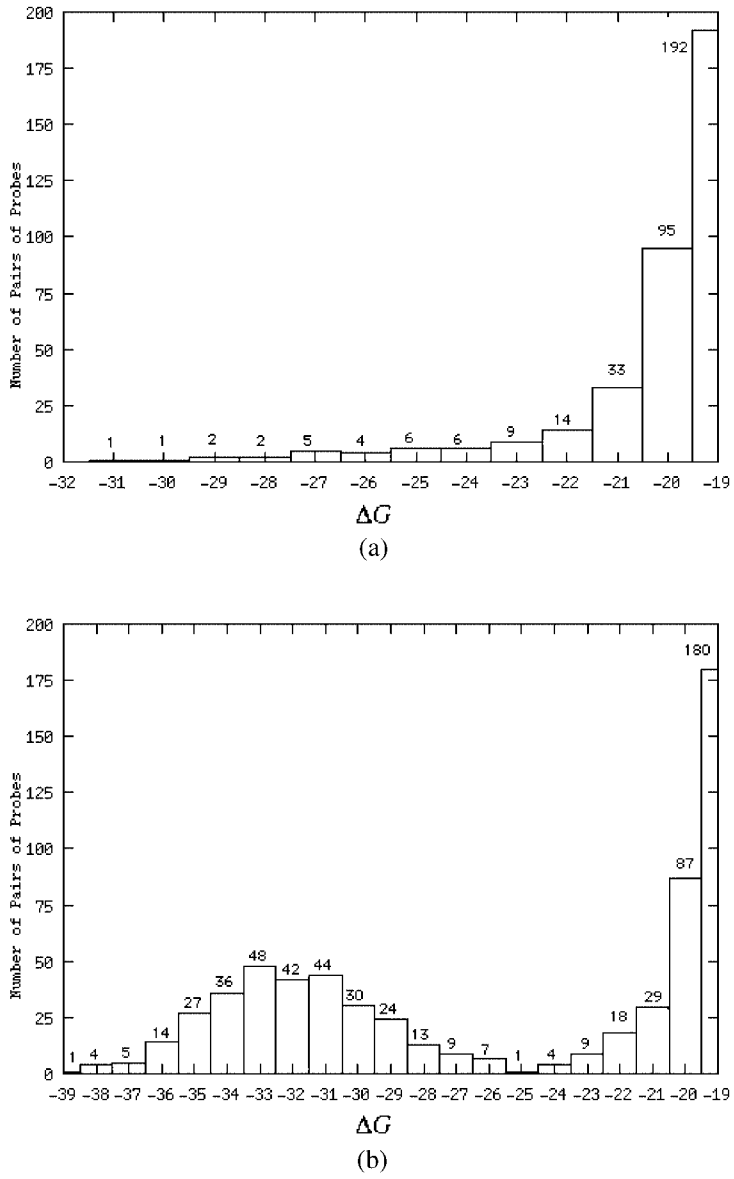


FIG. 1. Histogram (**a**) for cross-hybridization, depicts the number of probe pairs $p, p' \in P$, $p \neq p'$, where $\mathcal{X}(p, p') \leq \tau_c$ versus ΔG. Histogram (**b**) for dimerization, depicts the number of probe pairs $p, p' \in P$, $p \neq p'$, where $\mathcal{D}(p, p') \leq \tau_c$ versus ΔG.

Example 1 (Cross-hybridization). *For probe* $p = GAAAGCGGAACTGTTTCGGAGAAGG$ *in probe group U22029_f_at and probe* $p' = GAAAGCGGTACTGTTTCGGAGAAGG$ *in probe group M33317_f_at,* $\mathcal{X}(p, p') = -27.0336$. *For probe* $p = CCCTGCTGCTCATCGAGTCGTGGCT$ *in probe group J03071_cds3_f_at and probe* $p' = CCCTGCTGCTCATCCAGTCGTGGCT$ *in probe group J00148_cds2_f_at,* $\mathcal{X}(p, p') = -28.6136$. *In both of these examples, probes* p *and* p' *differ by exactly one base and belong to different probe groups.*

Example 2 (Dimerization). *An example of possible dimerization is that probe* $p = CGAAGCGGA$ $ATTCTCCATGCCCGAG$ *in probe group M24899_at and probe* $p' = CTCGGGCATGGAGAATT$ $CCGCTTCG$ *in probe group X72632_s_at have* $\mathcal{D}(p, p') = -32.4935$. *Note that probes* p *and* p' *are wc-complements.* AU1

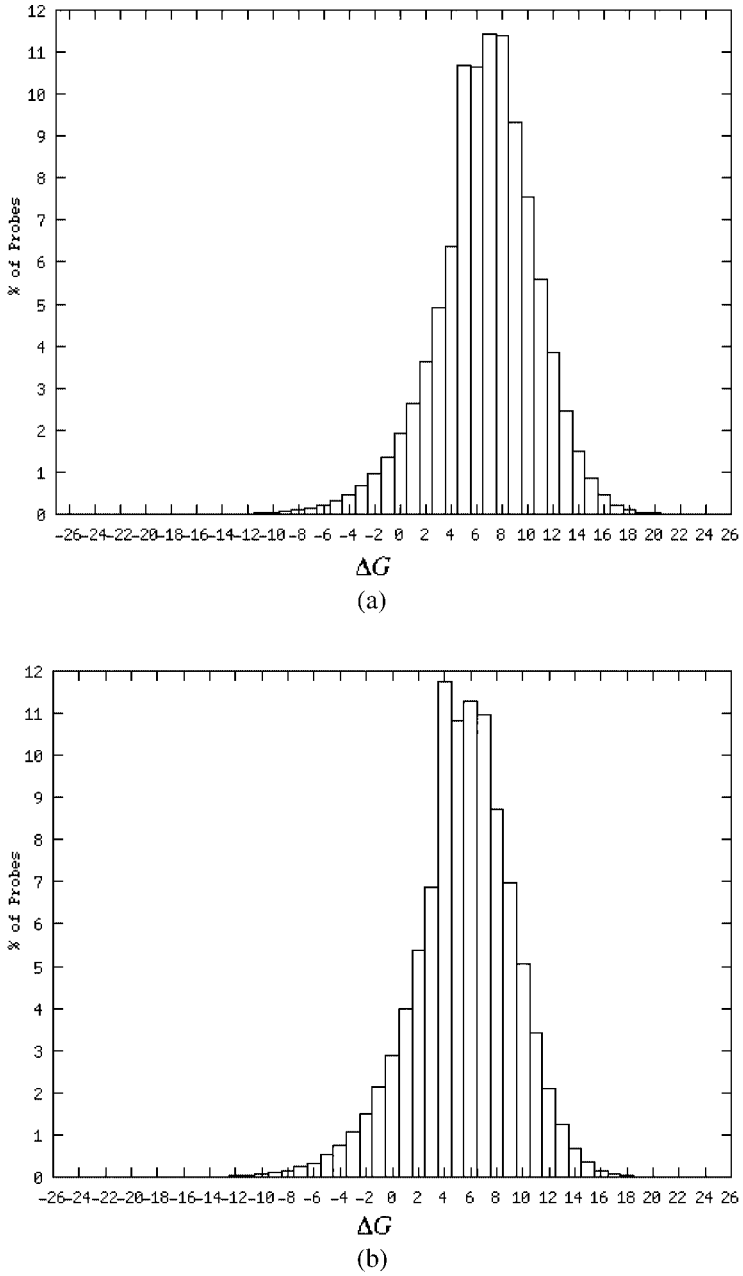


FIG. 2. For self-dimerization, histogram (**a**) depicts the distribution of ΔG estimates from $\mathcal{D}(p, p)$ for 10^8 randomly chosen $p \in \Sigma_{dna}^{25}$, and (**b**) depicts the distribution of ΔG estimates from $\mathcal{D}(p, p)$ for all probes $p \in P$.

Several additional examples of such probe pairs are also given below Fig. 3 of Smith and Hallett (2004).

Consider now the case of self-dimerization (dimerization between two copies of the same probe). We begin by comparing the distribution of the ΔG estimates for a set of 10^8 randomly chosen probes $p \in \Sigma_{dna}^{25}$ and the distribution of the ΔG estimates for $\mathcal{D}(p, p)$ for all $p \in P$. Figure 2 depicts these two histograms. F2 We find that the distribution of ΔG estimates $\mathcal{D}(p, p)$, for $p \in P$, is slightly more concentrated around its mean than the distribution of ΔG estimates for the random probe set. In particular, we find that the minimum ΔG measured by $\mathcal{D}(\cdot, \cdot)$ over all probes in P is -16.303 compared to -27.763 for the random set. Although there is no probe in set P such that $\mathcal{D}(p, p) \leq \tau_c$, the similarity of these two distributions is an indication that there do exist probes that have a high affinity to self-dimerize. Note, however, that

the minimum ΔG measured by $\mathcal{D}(\cdot,\cdot)$ over all probes in P is still much higher than the lowest possible such ΔG estimates.

Example 3 (Self-dimerization). *An example of a probe measured to have low estimates from $\mathcal{D}(p, p)$ in the case of self-dimerization is $p = TGTGTGGCGGTGACACCGTCACCCA$ with $\mathcal{D}(p, p) = -15.6435$. In this example, if two copies of p were to align with each other in opposing directions, they can hybridize with only four mismatches.*

The range of ΔG estimates for secondary structure formation of a probe differs from the distribution of ΔG estimates for hybridizations between a probe and its wc-complement. To examine ΔG estimates for secondary structure, we apply $\mathcal{S}(p)$ to a set of 10^8 randomly chosen probes $p \in \Sigma_{dna}^{25}$ and compare the distribution of the ΔG estimates with the distribution of ΔG estimates obtained from computing $\mathcal{S}(p)$, for all $p \in P$ from chip C. Figure 3 depicts these two histograms. As is the case with self-dimerization, **F3** the lowest ΔG estimate obtained from $\mathcal{S}(\cdot)$ over set P is much higher that the lowest ΔG measured by $\mathcal{S}(\cdot)$ over the random set of probes. We find that $-8.678 \leq \mathcal{S}(p') \leq 8.525$ for all probes $p' \in P$. This should be compared with $-14.579 \leq \mathcal{S}(p) \leq 11.335$ for the 10^8 randomly chosen probes $p \in \Sigma_{dna}^{25}$.

Example 4 (Secondary structure formation). *Examples of probes measured to have low $\mathcal{S}(\cdot)$ include $p = GCCACCACACTGGTGTGCTGGCTGT$ with $\mathcal{S}(p) = -8.67883$ and $p' = GCGAGGAAGC TTCCTCGCAACTTTG$ with $\mathcal{S}(p') = -7.36687$. It is the case that both p and p' can form a secondary structure where only very few base pairs are mismatched.*

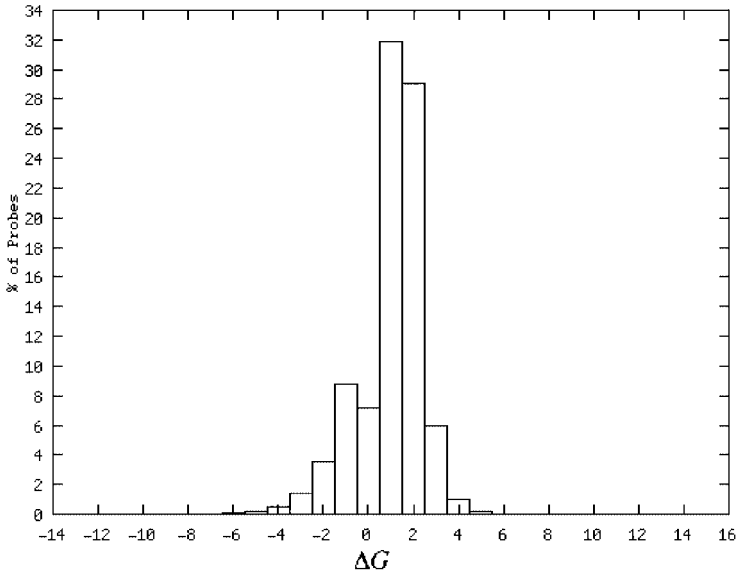
3.2. Analysis of the conflict graph

Let M be the conflict graph constructed from the HuGeneFL chip C as in Section 2.3. Table 1 displays **T1** the size of sets S, SD, X, D for various values of $\tau_s, \tau_{sd}, \tau_x$ and τ_d. Using Table 1 and the analysis of the previous subsection, our analysis is done w.r.t. the conflict graph M induced by $\tau_s = -6$, $\tau_{sd} = -14$, $\tau_x = -23$, and $\tau_d = -33$.

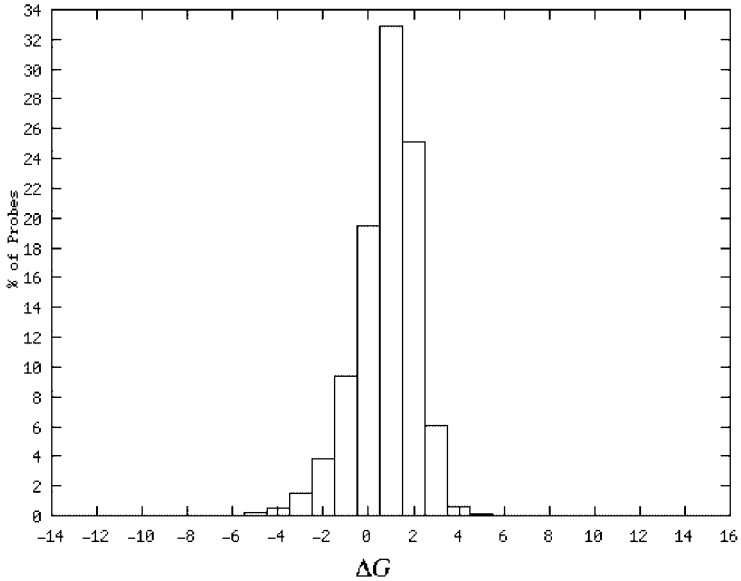
For these threshold values, let $S' = \{p \in V(G) : (p, p) \in S\}$ be the set of probes predicted to exhibit secondary structure, and $SD' = \{p \in V(G) : (p, p) \in SD\}$ be the set of probes predicted to exhibit self-dimerization. We find that $S' \cap SD' = \emptyset$, and therefore no single probe is predicted to have an affinity for both secondary structure formation and self-dimerization. We also find that every probe group P_g, $g \in G$ contains at most one probe in S' and one probe in SD'. For probe $p \in S'$ with matching tag t, the average over all ΔG estimates obtained from $\mathcal{DP}(p, t)$ is -33.67294. If we consider Fig. 7 of Smith and Hallett (2004), we see that these probe–tag pairs tend to have low ΔG estimates that range between -37.5535 and -29.1635. This indicates that they tend to be strong probe–tag pairs. In fact, the average over all ΔG estimates obtained from $\mathcal{DP}(p, t)$ for probes $p \in SD$ with matching tags t is also very low at -32.52438 with a range of -35.2135 to -28.6135 (Fig. 8 of Smith and Hallett [2004]).

Similarly, let D' and X' be the sets of probes predicted to exhibit dimerization and cross-hybridization, respectively. Consider a probe group P_g such that there exists a probe $p \in P_g$ and $p \in D'$. We find that any such probe group P_g contains between 3 and 10 probes in D'. The majority of probe groups P_g with a probe $p \in P_g$ such that $p \in X'$ do not contain additional probes in X'. However, a small number of probe groups do contain as many as six probes in X'. The probe–tag pairs for probes $p \in D'$ with matching tags t tend to have low ΔG estimates, as the values of $\mathcal{DP}(p, t)$ range between -38.2435 and -33.0435 with an average value of -34.5052 (Fig. 6 of Smith and Hallett [2004]). Although the average value of $\mathcal{DP}(p, t)$ for probes $p \in X'$ with matching tags t is very low and is close to the average value for dimerization at -31.6018, the range of values for X' ($-36.1735.. - 25.7035$) is much larger (Fig. 4 of Smith and Hallett [2004]).

We now consider vertices incident to cross-hybridization and dimerization edges. Let M_x be subgraph of M restricted to edges labeled x. Formally, $V(M_x) = V(M)$ and $E(M_x) = \{(p, p') \in E(M) : \kappa(p, p') = x\}$. It is the case that the majority of nonzero degree vertices in M_x have degree one, although the maximum degree of M_x is three. Let M_d be M restricted to edges labeled d and excluding self-loops, $V(M_d) = V(M)$ and $E(M_d) = \{(p, p') \in E(M) : \kappa(p, p') = d$ and $p \neq p'\}$. For the HuGeneFL chip, all the vertices in M_d have either degree zero or one. There are no vertices in $M_{xd} = M_x \cup M_d$ incident to both x and d labeled edges. Table 2 displays the number of vertices with degree i in M_x, M_d and $M_x d$. **T2**



(a)



(b)

FIG. 3. For self-hybridization, histrogram (**a**) depicts the percentage of ΔG estimates obtained from $\mathcal{S}(p)$ for 10^8 randomly chosen $p \in \Sigma_{dna}^{25}$, and histogram (**b**) depicts the percentage of ΔG estimates obtained from $\mathcal{S}(p)$ for all probes $p \in P$.

TABLE 1. SIZES OF EDGE SETS FROM THE CONFLICT GRAPH M INDUCED BY VARIOUS THRESHOLD VALUES IN KCAL/MOL FOR CHIP C

τ_s (kcal/mol)	$\lvert S\rvert$	τ_{sd} (kcal/mol)	$\lvert SD\rvert$	τ_x (kcal/mol)	$\lvert X\rvert$	τ_d (kcal/mol)	$\lvert D\rvert$
−7	5	−16	2	−30	3	−36	10
−6	33	−14	12	−23	27	−33	87
−5	92	−12	38	−20	83	−30	224
−4	318	−10	114	−18	458	−28	281
		−8	325				

TABLE 2. NUMBER OF VERTICES WITH DEGREE EXACTLY i AND AVERAGE ΔG FOR PROBES REPRESENTED BY VERTICES WITH DEGREE EXACTLY i FOR SUBGRAPHS M_x, M_d AND M_{xd}

Degree	$\lvert V(M_x)\rvert$	*Ave.* ΔG	$\lvert V(M_d)\rvert$	*Ave.* ΔG	$\lvert V(M_{xd})\rvert$	*Ave.* ΔG
0	127503	−29.140	127386	−29.134	127329	−29.137
1	53	−31.442	174	−34.50	227	−32.971
2	3	−29.376	0		3	−29.376
3	1	−29.953	0		1	−29.953

4. PATTERNS IN HYBRIDIZATION EXPERIMENTS

This section gives the formal definitions for the pattern of both nondegenerate probes and degenerate probes over a set of hybridization experiments. We use this framework to examine the sets of probes predicted to be degenerate and to experimentally examine the Affymetrix HuGeneFL chip in Section 5.

Let $H = \{H_1, \ldots, H_K\}$ be the set of hybridization experiments for a chip C. The output of a hybridization experiment H_i is simply an intensity value for every probe on chip C. The *intensity of probe p in hybridization H_j* is represented by $I_j(p) \in \mathbb{R}$. An estimate of the intensity for each target $g \in G$ is calculated from the members of the probe group of g, P_g. The *intensity of target $g \in G$ in experiment H_j* is represented by $I_j(g) \in \mathbb{R}$. For simplicity, we use an uncorrected quantitative measurement of the intensity of the target. For our purposes, $I_j(g) = \frac{\Sigma_{p \in P_g} I_j(p)}{|P_g|}$. Using the minimum and maximum intensity levels for a probe p (for a target g) over the set of hybridizations, the intensity measurements of a probe (of a target) for all hybridizations are scaled to the $(0 \ldots 1]$ interval. In a standard hybridization experiment, the expression level of a target is determined via a robust statistical method taking into account, for instance, discrimination of the probe–tag pair (Affymetrix, 2002; Li and Wong, 2001a, 2001b). We could also make use of these robust variants for calculating $I_j(g)$. These issues are discussed in greater detail by Smith and Hallett (2004) where this framework is integrated with the model from Li and Wong (2001a, 2001b).

At each experiment $H_i \in H$ and each target $g \in G$, the intensity of each probe $p \in P_g$ is *ranked* relative to the intensity of all remaining probes $P_g \setminus \{p\}$. For simplicity of exposition, we assume that all intensity measurements for a probe group are distinct.

Definition 2 (Rank). *The* rank *of a probe $p \in P_g$ in experiment H_j (written $\rho_j(p, g)$) is i iff there exist exactly $i - 1$ distinct elements $p_1, \ldots, p_{i-1} \in P_g \setminus \{p\}$ s.t. $I_j(p_k) < I_j(p)$, for $1 \le k \le i - 1$. When the probe group is clear from the context, we denote the rank simply as $\rho_j(p)$.*

We discretize the hybridization experiments into blocks according to $I_j(g)$ and use $b \in \mathbb{Z}$ to represent the desired number of blocks of the (0..1] interval.

Definition 3 (Block). *For a gene g in hybridization H_j, we say that* $I_j(g)$ is in block b' *iff* $\frac{b'-1}{b} < I_j(g) \le \frac{b'}{b}$.

The following definitions relate the rank of a probe p to the intensity of the target of p. We assume everywhere that the size of all probe groups is l.

Definition 4 (Occurrence). *We say that a probe p in hybridization H_j is a* rank i, block b' occurrence *iff $\rho_j(p) = i$ and $I_j(g)$ is in block b', where $p \in P_g$, $1 \le i \le l$ and $1 \le j \le K$.*

Definition 5 (Pairwise occurrence). *We say that a pair of probes (p, p') in hybridization H_j is a* rank i, block pair (b_1, b_2) occurrence *iff either*

(i) p is a rank i, block b_1 occurrence *and $I_j(g')$ is in block b_2, or*
(ii) p' is a rank i, block b_1 occurrence *and $I_j(g)$ is in block b_2,*

where $p \in P_g$, $p' \in P_{g'}$, $1 \le i \le l$ and $1 \le j \le K$.

We are interested in the number of times a set of probes is observed to have a specific rank over a set of hybridizations.

Definition 6 (Rank count vector). *For $P' \subseteq P$ and $H' \subseteq H$ let $y_{b'}^{P',H'}$ be the rank count vector where $y_{b'}^{P',H'}[i]$ is the number of rank i, block b' occurrences over all probes $p \in P'$ and all hybridizations $h \in H'$.*

Definition 7 (Pairwise rank count vector). *For $P' \subseteq P \times P$ and $H' \subseteq H$, let $y_{(b_1,b_2)}^{P',H'}$ be the rank count vector where $y_{(b_1,b_2)}^{P',H'}[i]$ is the number of rank i, block pair (b_1, b_2) occurrences over all probes pairs $(p, p') \in P'$ and all hybridizations $h \in H'$.*

The rank count vector (or pairwise rank count vector) describes the distribution of ranks for a set of probes over a set of hybridization experiments. However, we require a family of distributions that allows us to formally describe behavior such as "the ranks for a set of probes over a set of hybridizations tend to be uniformly distributed," or "the ranks for a set of probes tend to always be low." The *beta distribution* with parameters α and β turns out to be very useful for describing this family of distributions, since it is a very flexible, continuous distribution defined over a fixed range and it has a wide variety of shapes useful for describing any pattern of ranks.

The probability density function of the beta distribution, the *beta density function*, with parameters $\alpha, \beta > 0$ is defined as

$$f_{\alpha,\beta}(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, 0 < u < 1$$

where $\Gamma(\cdot)$ is the Gamma function generalizing the factorial expression for the natural numbers. We define $f_{\alpha,\beta}$ as the beta distribution with beta density function $f_{\alpha,\beta}(u)$. When $\alpha = \beta = 1$, $f_{\alpha,\beta}$ is the uniform distribution. When $\alpha \leq 1$ and β is large (and vice versa), $f_{\alpha,\beta}$ is an exponential distribution. We fit a beta distribution to a rank count vector by estimating parameters α and β from the rank count vector. The α and β parameters for a beta distribution can be estimated from sample x as follows

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right) \qquad \text{and} \qquad \hat{\beta} = (1-\bar{x})\left(\bar{x}\frac{1-\bar{x}}{s^2} - 1\right), \tag{1}$$

where $\bar{x}$ is the sample mean and s^2 is the unadjusted sample variance.

We also require a measure of how well a particular rank count vector fits a particular beta distribution. Towards this end, we define a discretization of the continuous beta distribution.

Definition 8 (Discretized probability vector). *The length l probability vector $\phi_{\alpha,\beta}$ is derived from $f_{\alpha,\beta}$ (with beta density function $f_{\alpha,\beta}(u)$) by*

$$\phi_{\alpha,\beta}[i] = \int_{(i-1)/l}^{i/l} f_{\alpha,\beta}(u)du, \text{ for } 1 \leq i \leq l.$$

We call $\phi_{\alpha,\beta}$ the discretized probability vector *of $f_{\alpha,\beta}$.*

Throughout this paper, we plot the discretized probability vectors to ease comparison between distributions. Figure 13 and Example 1 of Smith and Hallett (2004) depict the distribution function of beta distributions for a variety of parameters α and β and an example of the discretized probability vectors.

4.1. Estimating statistical significance

We use the following simple method for estimating the statistical significance of the patterns described below. For a probe p, we assume we know the rank of p within its probe group P_g over a set of k hybridization experiments. We model this with k independent identically distributed random variables $X_1, \ldots, X_k$ with state space $[1..l]$, $l = |P_g|$. Of course, hybridization experiments in practice may not

be independent; this is a simplifying assumption. The probability for each state is described by vector $\phi = \langle \phi_1, \ldots, \phi_l \rangle$, $\Sigma_i \phi_i = 1$. Let $Y = Y_1, \ldots, Y_l$ be random variables that count the number of times each of the l values (i.e., the l ranks) occur over $X_1, \ldots, X_k$. The probability that $Y[i] = y[i]$ for $1 \leq i \leq l$ is given by multinomial distribution formula

$$P_{Y,\phi}(y) = \frac{k!}{\prod_i (y[i]!)} \prod_i \phi[i]^{y[i]},$$

for a rank count vector y.

At various places throughout the paper, we have two discrete probability vectors $\theta = \langle \theta_1, \ldots, \theta_l \rangle$ and $\phi = \langle \phi_1, \ldots, \phi_l \rangle$. Here, θ is treated as the null hypothesis and typically represents a relevant background distribution of the rank of a probe over k hybridizations. The second vector ϕ represents the alternative hypothesis and typically represents a distribution of the rank of a probe over k hybridizations for probes known to exhibit a specific type of degenerate behavior (e.g., secondary structure, self-dimerization, cross-hybridization, dimerization). We ask how often we would expect it to be the case that $P_{Y,\phi}(y)$ is greater than $P_{Y,\theta}(y)$, for a rank count vector y generated randomly according to the distribution θ. Note that for our test it is always the case that $\Sigma_i y_i = k$.

More precisely, let $1_{Y,\phi,\theta}(y)$ be the indicator function defined as

$$1_{Y,\phi,\theta}(y) = \begin{cases} 1 & \text{if } P_{Y,\phi}(y) > P_{Y,\theta}(y) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where y is a rank count vector generated randomly according to distribution θ. We compute the mean of $1_{Y,\phi,\theta}$, $\bar{1}_{Y,\phi,\theta}$ by generating a set of r rank count vectors with $\Sigma_i y_i = k$ and computing $\Sigma_y 1_{Y,\phi,\theta}(y)/r$, for sufficiently large r. If $\bar{1}_{Y,\phi,\theta} \geq \epsilon$, then we say that the probability distributions represented by θ and ϕ are significantly different for ϵ and k.

4.2. Background distributions: Patterns of nondegenerate behavior

Throughout the following, let $C = \langle G, P, \{P_g : g \in G\}, T, \{T_g : g \in G\} \rangle$ represent the Affymetrix HuGeneFL chip and M be the conflict graph induced by C. Let $H = \{H_1, \ldots, H_K\}$ be our set of hybridization experiments.

We assume that the vast majority of probes of a chip exhibit nondegenerate behavior. It follows from this assumption that if we randomly select a large set of probes from P, then the aggregate distribution of ranks for this set of probes is a reasonable approximation to the distribution of ranks for nondegenerate probes. We call the distribution of ranks for nondegenerate probes the *background beta distribution* of ranks over the set of all probes P.

Let P' be a randomly chosen set of probes from P s.t. it is not the case that $p, p' \in P_g$ for some g, $p \neq p'$. Let $y_i^{P',H}$ be the rank count vector of block i over the set of all hybridizations H for $1 \leq i \leq 3$. The background beta distribution of ranks of probe set P for block i is defined by the beta distribution parameters $\widehat{\alpha}_i, \widehat{\beta}_i$ estimated from $y_i^{P',H}$ (Equations 1).

The discretized probability vector $\phi_{\widehat{\alpha}_i,\widehat{\beta}_i}$ is calculated from the resulting background beta distribution $f_{\widehat{\alpha}_i,\widehat{\beta}_i}$ for $1 \leq i \leq 3$ (Definition 8). These discretized probability vectors $\phi_{\widehat{\alpha}_i,\widehat{\beta}_i}$ are depicted in Fig. 4. As **F4** expected, these distributions suggest that the ranks of nondegenerate probes are uniformly distributed. If we consider a probe $p \in P'$ that we "knew" somehow was nondegenerate, then we might expect that the rank of p would be (close to) uniformly distributed over $[1 : |P_g|]$, where $p \in P_g$ for some g. Note also that the distributions do not vary greatly between the different blocks.

It turns out that we can do better than the simple uniform background beta distribution described above. In particular, we find that the distribution of ranks for a probe p over many hybridization experiments is dependent on the hybridization strength of p and its matching tag t. The discretized probability vectors calculated from the background beta distribution of probes for which $\mathcal{DP}(p,t)$ is low (little free energy; strong hybridization) depicted in Fig. 9 of Smith and Hallett (2004) shows that the higher ranks have much higher probability of occurring than do the low ranks. Therefore, if $\mathcal{DP}(p,t)$ is low, then a nondegenerate probe p should have a tendency to exhibit high ranks over a set of hybridization experiments. Otherwise,

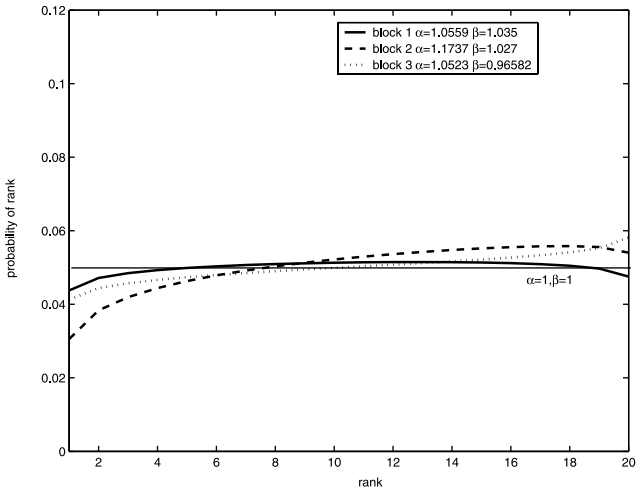


FIG. 4. Discretized probability vectors $\phi_{\widehat{\alpha}_i,\widehat{\beta}_i}$ calculated from $f_{\widehat{\alpha}_i,\widehat{\beta}_i}$ the estimated background beta distribution of the set of all probes P, for $b = 3$.

as depicted in Fig. 11 of Smith and Hallett (2004), the discretized probability vectors calculated from the background beta distribution of probes for which $\mathcal{DP}(p, t)$ is high (weak hybridization) show that a nondegenerate probe p should have a tendency to exhibit low ranks. These results confirm that ΔG is correlated with the pattern of ranks for nondegenerate probes. We also find that the rank of a nondegenerate probe does not vary greatly over different blocks (it is not a function of the intensity of its target). Figures 9 to 11 of Smith and Hallett (2004) depict this for the HuGeneFL chip and three blocks.

We now consider the background distribution of ranks for probes predicted to be degenerate. Let $S' = \{p \in V(G) : (p, p) \in S\}$ be the set of probes predicted to exhibit secondary structure. Let $\tau_{min,S}$ and $\tau_{max,S}$ specify the range of ΔG values calculated by $\mathcal{DP}(p, t)$ over each probe $p \in S'$ with matching tag t. Let $P'_S \subseteq P$ be the set of probes $p \in P$ s.t. $\tau_{min,S} \leq \mathcal{DP}(p, t) \leq \tau_{max,S}$. In other words, P'_S represents the subset of probe–tag pairs with a ΔG estimate similar to that for probes in S'. Using the set P'_S and Equations 1, we compute beta distribution parameters $\hat{\alpha}_S, \hat{\beta}_S$. The resulting discretized probability vector θ_S computed from $f_{\hat{\alpha}_S,\hat{\beta}_S}$ serves as the background distribution for secondary structure. We repeat this to determine background distributions θ_{SD}, θ_X, and θ_D for self-dimerization, cross-hybridization, and dimerization, respectively. These distributions are shown in Figs. 5 and 6. **F5 & F6**

4.3. Patterns for degenerate behavior

We now examine the pattern of ranks of probes predicted to be degenerate. We provide an intuitive hypothesis for the pattern of ranks of a probe with an affinity for each one of the four types of degeneracy. We justify these conjectures by showing that the rank count vectors of vertices incident to edges from the sets S, SD, X, D of conflict graph M do in fact follow these distributions. We use three blocks in the examination of single probe behavior (secondary structure and self-dimerization) and four blocks in the examination of pairwise behavior (cross-hybridization and dimerization). Ideally, we would like to use as many blocks as possible in order to clearly show the difference in intensity between blocks. However, when b is too large, some blocks are empty.

4.3.1. Secondary structure. Consider a target $g \in G$ with corresponding probe group $P_g = \{p, p_1, \ldots, p_{l-1}\}$ and suppose that it is known that p has a high affinity to form secondary structure. Furthermore, suppose that p is the only degenerate probe in P_g. We conjecture that the intensity of p w.r.t. P_g will follow two principles. First, if the target g is highly expressed in hybridization experiment H_j, the intensity of probes $I_j(p_i)$, $1 \leq i \leq l - 1$, will be higher than the intensity of p, $I_j(p)$. This is due to the fact that tag t is not hybridizing with p during the experiment at the same rate as other nondegenerate tags hybridize with members of P_g. Therefore, the rank of p in this experiment, $\rho_j(p)$, is expected to be very

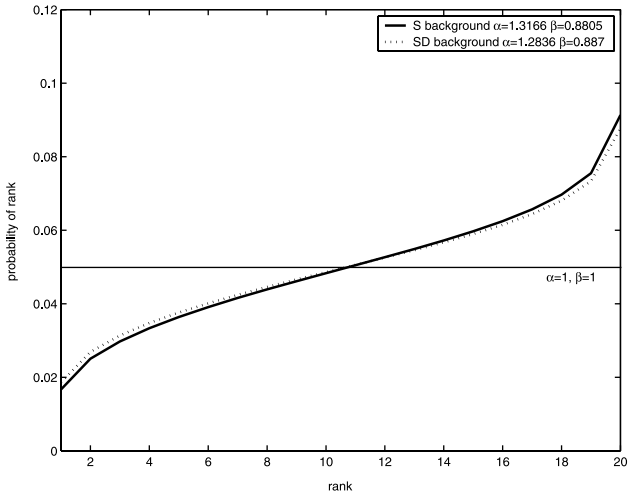


FIG. 5. Discretized probability vectors θ_S and θ_{SD} calculated from the estimated background beta distribution for the set of probes $P'_S \subseteq P$ such that for $p \in P'_S$ with matching tag t, $\tau_{min,S} \leq \mathcal{DP}(p,t) \leq \tau_{max,S}$ and from the estimated background beta distribution for the set of probes $P'_{SD} \subseteq P$ such that for $p \in P'_{SD}$ with matching tag t, $\tau_{min,SD} \leq \mathcal{DP}(p,t) \leq \tau_{max,SD}$.

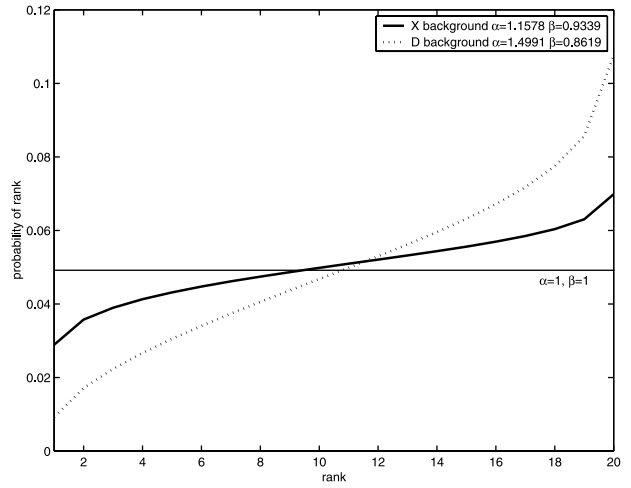


FIG. 6. Discretized probability vectors θ_X and θ_D calculated from the estimated background beta distribution for the set of probes $P'_X \subseteq P$ such that for $p \in P'_X$ with matching tag t, $\tau_{min,X} \leq \mathcal{DP}(p,t) \leq \tau_{max,X}$ and from the estimated background beta distribution for the set of probes $P'_D \subseteq P$ such that for $p \in P'_D$ with matching tag t, $\tau_{min,D} \leq \mathcal{DP}(p,t) \leq \tau_{max,D}$.

low. Second, if the target g is lowly expressed in hybridization experiment H_j, the difference in intensity between members of P_g will be small.

Let $S' = \{p \in V(G) : (p,p) \in S\}$ be the set of probes predicted to exhibit secondary structure. Let $G' \subseteq G$ be the set of targets induced by S'. We bin the intensity values for all targets $g \in G'$ into b blocks (Definition 3). For S' and hybridization set H, let $y_i = y_i^{S',H}$ be the rank count vector as defined in Definition 6, for each i, $1 \leq i \leq b$. Using Equation 1, we compute parameters $\widehat{\alpha}_i$, $\widehat{\beta}_i$ for a beta distribution.

The resulting discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ calculated from $f_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \leq i \leq b = 3$ depicted in Fig. 7 reaffirm our intuition. When $i = 1$ (the targets are lowly expressed), the beta distribution is near uniform $\alpha = 1.392$, $\beta = 1.4247$ (the ranks are almost uniformly distributed). When $i = 2$ (the targets are moderately expressed), the beta distribution is now nonuniform $\alpha = 1.037$, $\beta = 1.3645$ (the ranks tend to

F7

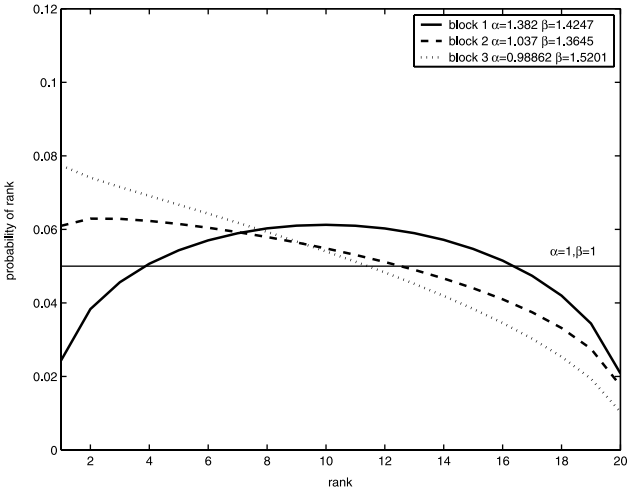


FIG. 7. Estimated discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \le i \le b$ for the set of probes S' predicted to have an affinity to form secondary structure for $\tau_s = -6$, $b = 3$.

be lower). Finally, when $i = 3$ (the targets are highly expressed), the beta distribution has negative slope $\alpha = 0.98862$, $\beta = 1.5201$ (the ranks tend to be extremely low and almost no high ranking probes). More formally, $\phi_{\widehat{\alpha}_1,\widehat{\beta}_1}$ has a nearly uniform distribution, and $\phi_{\widehat{\alpha}_i,\widehat{\beta}_i}$ has a distribution that approaches exponential with $\phi_{\widehat{\alpha}_i,\widehat{\beta}_i}[j] > \phi_{\widehat{\alpha}_i,\widehat{\beta}_i}[j+1]$, as i approaches b.

Recall the discussion concerning the background beta distribution for probes with hybridization strength within the ΔG range of probes in S' in Section 4.2. If the probes in S' were nondegenerate, they would be expected to follow the background beta distribution depicted by θ_S in Fig. 5. In fact, the curves for $i = 2$ and $i = 3$ in Fig. 7 contradict this.

Let θ_S be the discretized probability vector for the background beta distribution (here, $\alpha = 1.3166$ and $\beta = 0.8805$) for probes with hybridization strength within the ΔG range of probes in S' as described in Section 4.2. We estimate the mean $\bar{1}_{Y,\phi_i,\theta_S}$ as specified in Equation 2 for θ and $\phi_i = \phi_{\widehat{\alpha}_i,\widehat{\beta}_i}$ with a set of $r = 10{,}000$ rank count vectors and with the same sample size as rank count vector y_i, for $1 \le i \le b$. In fact, we find that $\bar{1}_{Y,\phi_i,\theta_S} = 0.0$ for all values of i. We conclude that each of the probability distributions represented by θ_S and ϕ_i are significantly different.

4.3.2. Self-dimerization. We conjecture that probes with an affinity to self-dimerize exhibit the same pattern of ranks as the pattern used for secondary structure. We test the conjecture by examining the set of probes $SD' = \{p \in V(G) : (p, p) \in SD\}$ predicted to exhibit self-dimerization. Let $y_i = y_i^{SD',H}$ be the observed count vector as defined in Definition 6 for each i, $1 \le i \le b$. As with secondary structure, we use Equation 1 to compute estimate parameters $\widehat{\alpha}_i$, $\widehat{\beta}_i$ for a beta distribution.

Figure 8 depicts the discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ calculated from the resulting $f_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \le i \le 3$. These distributions reaffirm our intuition of the pattern of ranks of self-dimerizing probes, although the evidence is not as strong as evidence for secondary structure. Note that when $i = 3$ (targets are more highly expressed), the number of high ranks is lower than when $i = 1$ (targets are more lowly expressed). Furthermore, the probability of low ranks is higher when $i = 3$ than when $i = 1$. **F8**

Additional evidence supporting the conjecture that self-dimerizing probes are behaving according to our pattern is obtained by considering the background beta distribution θ_{SD} for probes with hybridization strength within the ΔG range of probes in SD' as discussed in Section 4.2 (here $\alpha = 1.2836$ and $\beta = 0.887$). We estimate the expected value $\bar{1}_{Y,\phi_i,\theta_{SD}}$ as specified in Equation 2 for θ_{SD}, and $\phi_i = \phi_{\widehat{\alpha}_i,\widehat{\beta}_i}$, for $1 \le i \le b$. For these estimates, we use a set of $r = 10{,}000$ rank count vectors and with the same sample size as rank count vector y_i, for $1 \le i \le b$. Here we find that $\bar{1}_{Y,\phi_1,\theta_{SD}} = 0.013$, $\bar{1}_{Y,\phi_2,\theta_{SD}} = 0.003$, and $\bar{1}_{Y,\phi_3,\theta_{SD}} = 0.133$. Therefore, we conclude that the the probability distributions represented by θ_{SD} and ϕ_i are significantly different. The sample sizes and values of $\bar{1}_{Y,\phi_i,\theta_{SD}}$ are available in Table 4 of Smith and Hallett (2004).

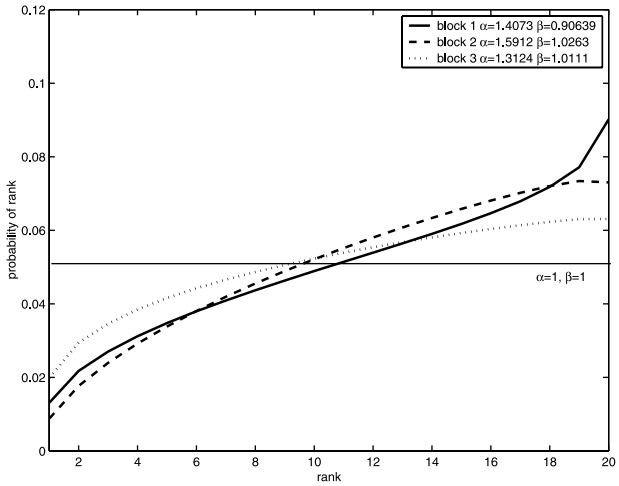


FIG. 8. Estimated discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \le i \le b$ for the probes belonging to self-dimerization set SD' for $\tau_{sd} = -14$, $b = 3$.

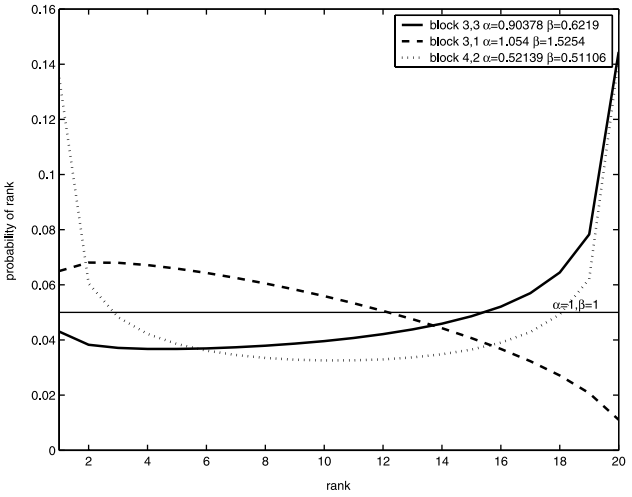
4.3.3. Cross-hybridization. Consider two distinct targets $g, g' \in G$ with corresponding probe groups $P_g = \{p, p_1, \ldots, p_{l-1}\}$, and $P_{g'} = \{p', p'_1, \ldots, p'_{l-1}\}$, and suppose that the tags t, t' have high affinities to cross-hybridize with p' and p, respectively. We say that the probe p gains tags from p', as some of the t' tags will not hybridize with p' but with p. Alternatively, the probe p loses tags to p', as some of the t tags will hybridize with p' but not with p. If g is lowly expressed and g' is highly expressed in hybridization H_j, then probe p will gain tags from p' but p' is not likely to gain tags from p. Therefore, p is expected to have a high rank w.r.t. the other elements of P_g, and p' is expected to have low rank w.r.t. the other elements of $P_{g'}$. Similarly, if g' is lowly expressed and g is highly expressed in hybridization H_j, then p' is expected to have a high rank w.r.t. the other elements of P_g, and p is expected to have low rank w.r.t. the other elements of $P_{g'}$. If both targets are equally expressed, then $\rho_j(p)$ and $\rho_j(p')$ are both expected to behave as the ranks of nondegenerate probes.

Let $X' = \{p, p' : (p, p') \in X\}$ be the set of probe pairs predicted to exhibit cross-hybridization. We bin the intensity values for each of the targets g into b blocks (Definition 3). For X' and hybridization set H, let $y_{(i,j)} = y_{(i,j)}^{X',H}$ be the observed pairwise rank count vector for each block pair as defined in Definition 7. Using Equation (1), we compute estimate parameters $\widehat{\alpha}_{i,j}, \widehat{\beta}_{i,j}$ for a beta distribution, $1 \le i, j \le b$. Figure 9 depicts the a subset of discretized probability vectors $\phi_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$ calculated from estimated beta distributions $f_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$, for $1 \le i, j \le b$ where $b = 4$. The discretized probability vector $\phi_{\widehat{\alpha}_{i,j},\widehat{\beta}_{i,j}}$ is labeled by *block* i, j. **F9**

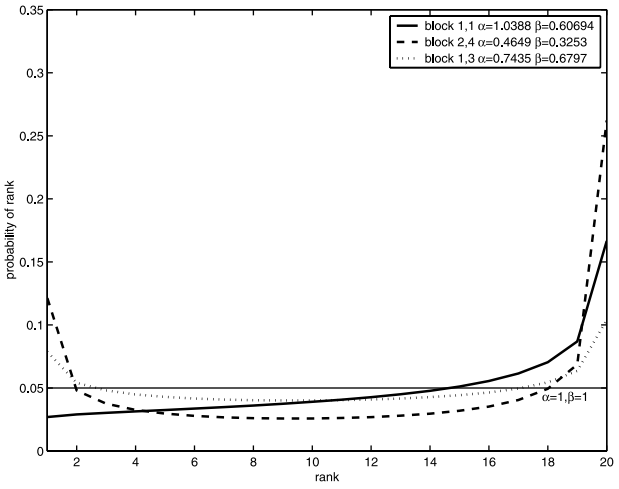
We argue that discretized probability vectors confirm that probes predicted to exhibit cross-hybridization have rank patterns that follow the pattern described above. Figure 9(a) depicts block pairs i, j where i is high ($i = 3$ or $i = 4$) and the value of j varies. That is, targets g are highly expressed, and targets g' have different expression levels.

Block pairs (3, 1) (and (4, 2)) in Fig. 9 are sets of hybridizations where p is expected to lose tags to p', since g is highly expressed and g' is lowly expressed. When we compare the solid curve for block 3, 3 to the bold dashed curve for block 3, 1, we see a strong difference in the distribution of high ranks that is in accordance with the intuition described above. Also in accordance with the pattern is the fact that there is a higher number of low ranks for probes for block 3, 1. **AU2**

Conversely, Fig. 9(b) depicts block pairs when i is fixed at a low value $i = 1$ or $i = 2$ and the value of j varies. The pattern for cross-hybridization is not confirmed for low ranks: in this case, the number of low ranks for blocks 1, 3 and 2, 4 is higher than the number of low ranks for block 1, 1. The pattern for cross-hybridization is also not confirmed for high ranks: in this case, the number of high ranks for block 1, 3 is lower than the number of high ranks for block 1, 1. If instead we compare blocks 2, 4 and 1, 1, this pattern is verified. We note, however, that the evidence for cross-hybridization is weaker than for the other types of degenerate behavior. The estimated parameters $\widehat{\alpha}_{i,j}, \widehat{\beta}_{i,j}$, $1 \le i, j \le b$ for all b^2 block pairs are included in Table 1 of Smith and Hallett (2004).



(a)

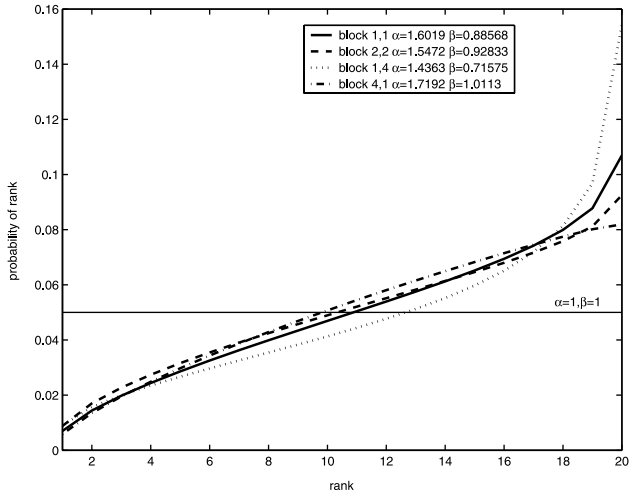


(b)

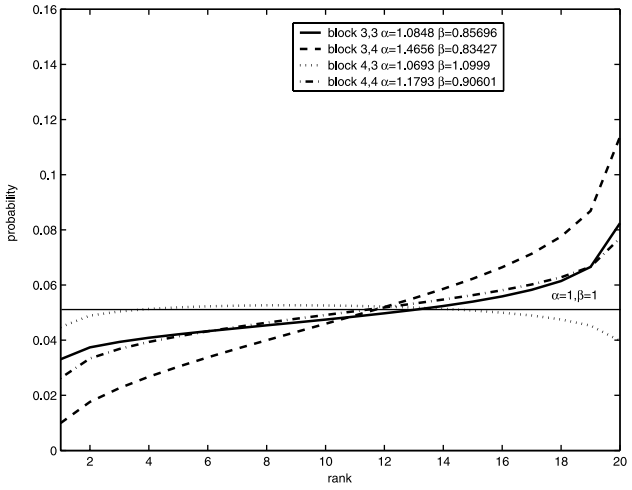
FIG. 9. Estimated discretized probability vectors $\phi_{\hat{\alpha},\hat{\beta}}$ for the set of probe pairs X' predicted to have an affinity to cross-hybridize with $\tau_X = -23$ and $b = 4$. (**a**) Block pairs (values of i, j) when i is fixed at a high value ($i \approx b$). (**b**) Block pairs when i is fixed at a low value ($i \approx 1$).

Additional evidence supporting the conjecture that cross-hybridizing probes are behaving according to our pattern is obtained by estimating the expected value $\bar{1}_{Y,\phi_{i,j},\theta_X}$ as specified in Equation 2, where θ_X is the discretized probability vector for the background distribution for cross-hybridizing probes and $\phi_{i,j} = \phi_{\widehat{\alpha}_{i,j},\widehat{\beta}_{i,j}}$, for $1 \leq i, j \leq b$. For these estimates, we use a set of $r = 10{,}000$ rank count vectors and with the same sample size as rank count vector $y_{(i,j)}$, for $1 \leq i \leq b$. The maximum value over all $\bar{1}_{Y,\phi_{i,j},\theta_X}$ is 0.137; therefore, we conclude that the the probability distributions represented by θ_X and $\phi_{i,j}$ are significantly different. The complete set of sample sizes and values of $\bar{1}_{Y,\phi_{i,j},\theta_X}$ are available in Table 5 of Smith and Hallett (2004).

4.3.4. Dimerization. Consider two distinct targets $g, g' \in G$ with corresponding probe groups $P_g = \{p, p_1, \ldots, p_{l-1}\}$ and $P_{g'} = \{p', p'_1, \ldots, p'_{l-1}\}$ and suppose that the corresponding tags t, t' have high a affinity to dimerize with each other. If both g and g' are highly expressed in hybridization H_j, then, since both t and t' are present in the sample, both p and p' will have fewer than expected tags hybridize with them. Therefore, p and p' are expected to have low ranks w.r.t. the other elements of P_g and $P_{g'}$. If it



(a)



(b)

FIG. 10. Estimated discretized probability vectors $\phi_{\hat{\alpha},\hat{\beta}}$ for the set of probe pairs D' predicted to have an affinity to dimerize with $\tau_d = -33$ and $b = 4$. (**a**) Block pairs where dimerization does not occur. (**b**) Block pairs where dimerization occurs.

is the case that (i) neither g nor g' is highly expressed or (ii) exactly one of g or g' is highly expressed but the other is not expressed, then, since only one of t or t' is present, the number of tags hybridizing to their respective probe will be as though no degeneracy existed. Therefore, the ranks $\rho_j(p)$ and $\rho_j(p')$ are both expected to behave as the ranks of nondegenerate probes.

Let $D' = \{p, p' : (p, p') \in D\}$ be probes predicted to exhibit dimerization. We bin the intensity values for each of the targets g into b blocks (Definition 3). Let $y_{(i,j)} = y_{(i,j)}^{D',H}$ be the observed count vector for each block pair as defined in Definition 7. Using Equations (1), we compute estimate parameters $\widehat{\alpha}_{(i,j)}, \widehat{\beta}_{(i,j)}$ for a beta distribution, $1 \le i, j \le b$. Figure 10 depicts a subset of discretized probability vectors $\phi_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$ calculated from estimated beta distributions $f_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$ for $1 \le i, j \le b = 4$. Discretized probability vector $\phi_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$ is labeled *block* i, j. **F10**

Figure 10(a) depicts block pairs where at least one of i or j is approximately equal to 1. Here we do not expect dimerization to effect the rank of the probes. Figure 10(b), however, depicts block pairs where both i and j are approximately equal to b. Here, we expect dimerization to raise the probability of low ranks and to decrease the probability of high ranks. It is important to recall that the probes in set D' tend

to have low ΔG estimates (Section 4.2) and are therefore expected to have rank distributions similar to the background beta distribution depicted by discretized probability vectors in Fig. 6. With the exception of block 3, 4, the distributions depicted in Fig. 10(b) show that the ranks are not following the background beta distribution and are consistent with the pattern for dimerization. Figures 10(a) and (b) together do give evidence that this behavior is occurring.

For both cross-hybridization and dimerization degeneracies, we give our results for $b = 4$. Other reasonable values of b gave similar results. The estimated parameters $\widehat{\alpha}_{i,j}$, $\widehat{\beta}_{i,j}$, $1 \le i, j \le b$ for all b^2 block pairs are available in Table 2 of Smith and Hallett (2004).

We again determine the difference between the probability distributions represented by θ_D, the discretized probability vector for the background beta distribution for probes predicted to dimerize, and the probability distributions estimated from D' shown by discretized probability vectors in Fig. 10. We find that the maximum value over all $\bar{1}_{Y,\phi_{i,j},\theta_D}$ is 0.405, where $\phi_{i,j} = \phi_{\widehat{\alpha}_{i,j},\widehat{\beta}_{i,j}}$, for $1 \le i, j \le b$. The high values of $\bar{1}_{Y,\phi_{i,j},\theta_D}$ occur in block pairs when dimerization should not be occurring and the probes are expected to be behaving as nondegenerate probes. With the exception of block pair (3, 4), we find that $\bar{1}_{Y,\phi_{i,j},\theta_D} = 0.0$ in block pairs when dimerization is expected to occur. These values for $\bar{1}_{Y,\phi_{i,j},\theta_D}$ over all block pairs (i, j), $1 \le i, j, \le b$ confirm that the distribution of ranks of probes in D' and the background beta distribution are very different when dimerization is expected to occur and are more similar when dimerization is not expected to occur. The complete set of sample sizes and values of $\bar{1}_{Y,\phi_{i,j},\theta_D}$ are available in Table 6 of Smith and Hallett (2004).

5. IDENTIFYING DEGENERATE PROBES

Consider any one of the probability vectors $\phi_{\alpha,\beta} = \phi = \langle \phi_1, \ldots, \phi_l \rangle$ derived in Section 4.2 (nondegenerate behavior), or Sections 4.3.1, 4.3.2, 4.3.3, or 4.3.4 (degenerate behaviors). Here ϕ_i is the probability that a probe p with an affinity for the particular nondegenerate or degenerate behavior is a rank i occurrence in a hybridization, $1 \le i \le l$. We use this probability vector ϕ to answer the following natural question: what is the probability of a rank count vector $y = \langle y_1, \ldots, y_l \rangle$ computed from a set of hybridization experiments, given that p is a probe with rank pattern described by ϕ? If ϕ corresponds to the probability vector for nondegenerate probes, then the above question asks for the probability of the rank counter vector y when p is assumed to be a nondegenerate probe. Otherwise, if ϕ is a probability vector for any of the degenerate behaviors, then we ask for the probability of the rank count vector when p is a degenerate probe.

The remainder of this section develops a set of support functions that determine whether a rank count vector is more likely to be distributed according to a pattern of degenerate behavior or more likely to be distributed according to a pattern for nondegenerate behavior. Such functions are important, since they would allow us to "learn" suspect probe–tag pairs from hybridization data (independently or together with theoretical models for hybridization). As the number of hybridization experiments increases for a particular chip, our ability to estimate the probability that a probe has an affinity for a particular type of degenerate behavior increases. This allows us to weight the intensity of such probes accordingly in the analysis of data from future experiments.

5.1. Support functions

5.1.1. Secondary structure. The secondary structure support function $\hat{\mathcal{S}}$ is the sum, over all blocks, of the log ratio of the probability of seeing the rank count vector given that the probe is prone to secondary structure and of the probability of seeing the rank count vector given that the probe is nondegenerate. Let ϕ_i, $1 \le i \le b$, be the discretized probability vector of $f_{\widehat{\alpha}_i,\widehat{\beta}_i}$ from Section 4.3.1. Let θ be the discretized probability vector for nondegenerate probes with ΔG estimates in the same ΔG range as $\mathcal{DP}(p, t)$ for probe p with matching tag t as discussed in Section 4.2. Here, θ is the discretized probability vector for the background beta distribution for p. Given the collection of count vectors $y = \langle y_1 \ldots y_b \rangle$ for p obtained from the set of hybrizations H, let

$$\widehat{\mathcal{S}}(p) = \sum_{i=1}^{b} log\left(\frac{P_{Y_i,\phi_i}(y_i)}{P_{Y_i,\theta}(y_i)}\right).$$

5.1.2. Self-dimerization. The self-dimerization support function $\widehat{\mathcal{SD}}(\cdot)$ is defined as the sum of the log ratio of the probability a rank count vector is seen, assuming the probe is prone to self-dimerization, and of the probability the probe is nondegenerate. Let ϕ_i, $1 \leq i \leq b$, be the discretized probability vector for $f_{\widehat{\alpha}_i,\widehat{\beta}_i}$. As with secondary structure, let θ be the discretized probability vector for the background beta distribution of ranks for probe p. Given the collection of count vectors $y = \langle y_1 \ldots y_b \rangle$ for probe p obtained from the set of hybrizations H, let

$$\widehat{\mathcal{SD}}(p) = \sum_{i=1}^{b} log\left(\frac{P_{Y_i,\phi_i}(y_i)}{P_{Y_i,\theta}(y_i)}\right).$$

5.1.3. Cross-hybridization. Let $\phi_{i,j}$ be the discretized probability vector $f_{\widehat{\alpha}_{(i,j)},\widehat{\beta}_{(i,j)}}$ for cross-hybridization from Section 4.3.3 for $1 \leq i, j \leq b$. Let θ be the discretized probability vector for nondegenerate probes with ΔG estimates in the same ΔG range as $\mathcal{DP}(p,t)$ and $\mathcal{DP}(p',t')$ for probes p, p' with matching tags t, t' as discussed in Section 4.2. Here, θ is the discretized probability vector for the background beta distribution for p and p'. We calculate the support $\widehat{\mathcal{X}}$ that (p, p') is a probe pair exhibiting cross-hybridization as follows:

$$\widehat{\mathcal{X}}(p, p') = \sum_{1 \leq i,j \leq b} log\left(\frac{P_{Y_{(i,j)},\phi(i,j)}(y_{(i,j)})}{P_{Y_{(i,j)},\theta}(y_{(i,j)})}\right).$$

5.1.4. Dimerization. Let $\phi_{(i,j)}$ be the discretized probability vector $f_{\widehat{\alpha}_{(i,j)},\widehat{\beta}_{(i,j)}}$ for dimerization from Section 4.3.4 for $1 \leq i, j \leq b$. Let θ be the discretized probability vector for the background beta distribution of probes p and p'. We compute the b^2 rank count vectors and denote these by $y = y_{(1,1)} \ldots y_{(b,b)}$. The dimerization support function $\widehat{\mathcal{D}}$ for a probe pair (p, p') is defined as follows:

$$\widehat{\mathcal{D}}(p, p') = \sum_{1 \leq i,j \leq b} log\left(\frac{P_{Y_{(i,j)},\phi_{(i,j)}}(y_{(i,j)})}{P_{Y_{(i,j)},\theta}(y_{(i,j)})}\right).$$

5.2. Experimental testing of support functions

To test how well these four support functions discriminate between nondegenerate and degenerate probes, the support scores for each set of degenerate probes (predicted by the theoretical models of hybridization) are compared with the support scores for a large randomly chosen set of probes. In total, we used 126 hybridizations from three different laboratories for the HuGeneFL chip. It is expected that the experimental support functions should assign large support values to probes (or probe pairs) predicted to be degenerate in the conflict graph M (probe sets S', SD', X', D'). However, Fig. 11 of this paper and Figs. 15, 16, and 17 of Smith and Hallett (2004) indicate that the support functions do not discriminate very well, since the support values measured for these probes (and probes pairs) appear random. Furthermore, the mean support score for *each* of the four support functions is significantly above zero. Several conclusions are possible from this. This may indicate that a large number of probes are degenerate. This seems, however, unlikely. It seems more likely that the support scores are calculated from simply too few hybridizations (i.e., $|H|$ is not sufficiently large). **F11**

We conjecture that our initial experiments here yield poor results due to the limited amount of data we used. The experimental support functions did not succeed in finding individual probes (or probe pairs) at a given block (or block pair), as there were simply too few hybridizations within the block (or block pair) to ensure that each of the element of the rank count vector has a sufficiently large frequency. One need consider that H contains only 126 hybridization experiments. For a probe p, this set of hybridizations is partitioned into $b = 3$ blocks. Some partitions had as few as five hybridization experiments.

We experimented with several other log-likelihood goodness-of-fit tests including adding pseudo-counts to the rank count vectors with these support functions. None of the alternative formulations resulted in significantly better results. For a chi-square test to succeed, it is best that each element of the rank count vector (i.e., each possible rank) be ≥ 5 (Moore *et al.*, 1995). For $|H| = 126$ and $b = 3$, most of the 20 individual elements of the rank count vectors have magnitude ≤ 5, and many are in fact 0. Using this lower

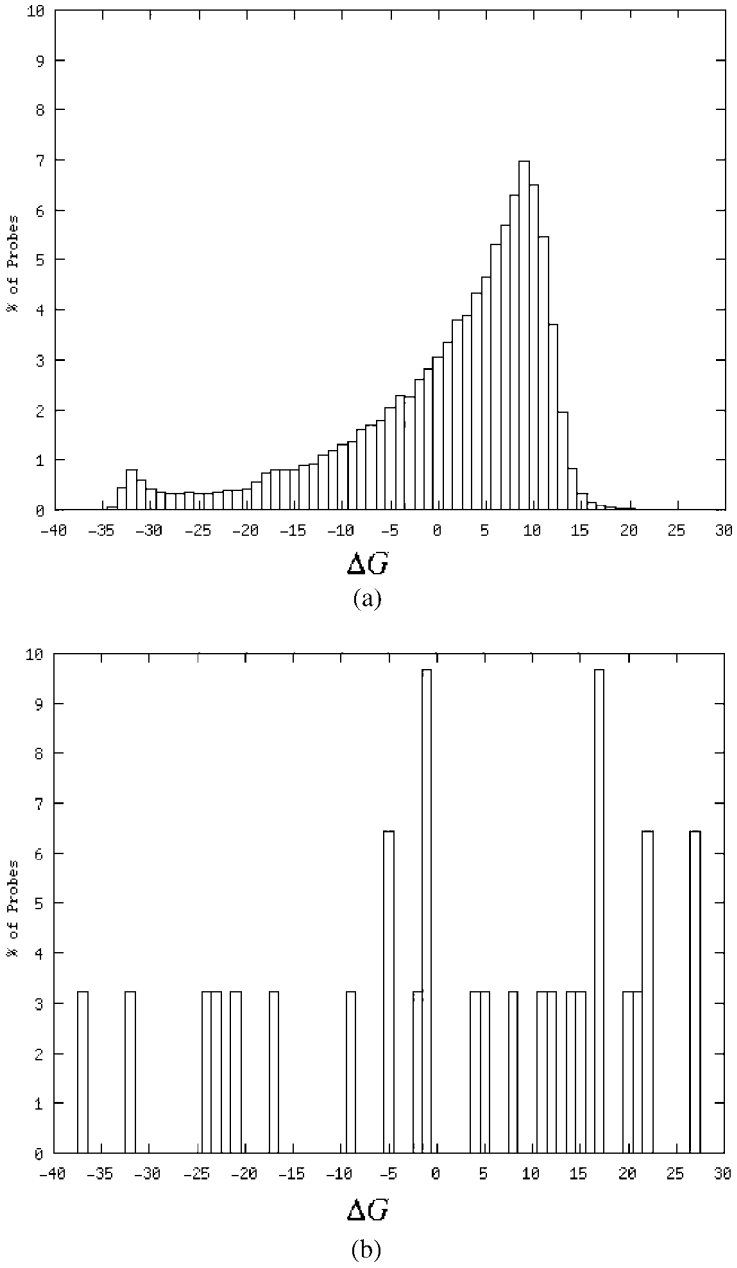


FIG. 11. Histograms of secondary structure formation support values measured by $\widehat{\mathcal{S}}$ for (**a**) the set of probes P and (**b**) the set of probes S'.

bound on the magnitude of each element of the vector, it is possible to estimate the minimum required size of the hybridization set H. Essentially, in our estimated probability distributions of degenerate probes and nondegenerate probes, the probability of a rank may be as low as 0.01. Therefore, some 500 hybridizations per block would be a very liberal estimate on the number of hybridizations required. Recall that the entire set of hybridizations must be partitioned into b blocks (or b^2 block pairs). For $b = 3$, we would require some 1,500 hybridizations for secondary structure and self-dimerization and some $1{,}500 \cdot 3^2$ for the pairwise tests of cross-hybridization and dimerization.

We attain more conservative estimates for the required number of hybridizations by calculating the mean of the indicator function $\bar{1}_{Y,\phi,\theta}(y)$ from Equation 2 with increasing sample sizes k. We find that $\bar{1}_{Y,\phi,\theta}(y)$ converges on ϵ, the expected value of $\bar{1}_{Y,\phi,\theta}(y)$, when $k \approx 300$. Therefore, we can determine the number

of hybridization experiments required to ensure that each rank count vector is generated from a sample size of at least 300. Such data now exists for some Affymetrix GeneChips, and our software has been designed to handle this magnitude of data.

6. AFFYMETRIX DISCRIMINATION

This section incorporates Affymetrix's *discrimination* into our framework, and we investigate the relationships between it and free-energy calculations. The discrimination property measures the ability of the intensity for a probe to represent the true amount of target mRNA transcripts in the sample. For each probe p in an Affymetrix GeneChip, there exists a so-called *mismatch (mm-) probe* p' (p is referred to as the *perfect match [pm-] probe*). The mm-probe p' differs from p by exactly one base pair (the base in the middle of the olignonucleotide). A hybridization experiment $H_j \in H$ returns both a measure of the intensity of probe $p \in P$, $I_j(p)$ and a measure of the intensity of the mm-probe p', written $\overline{I_j(p)}$.

The intensity analysis of probes in Affymetrix GeneChips is performed by one of several statistical detection algorithms. In essence, the detection algorithms combine "votes" from each probe in a probe group to assign a call of *present*, *marginal*, or *absent* to the target of the probe group (Affymetrix, 2002). The vote of each probe at hybridization $H_j \in H$ is simply the discrimination score $R_j(p)$ defined as follows:

$$R_j(p) = \frac{I_j(p) - \overline{I_j(p)}}{I_j(p) + \overline{I_j(p)}}.$$

The detection algorithm calculates a *detection p-value* according to the discrimination score of each probe p in the probe group P_g. If the majority of probes in P_g have $R_j(p) \approx 1$, then the detection p-value is significant and the transcript is likely assigned a call of present. Otherwise, if the majority of probes $p \in P_j$ have $R_j(p)$ near or below zero, then the detection p-value is not significant and the transcript is assigned a call of absent. If the detection p-value is above or below user-defined thresholds, then the transcript receives a marginal call.

Let $R_j(p)$ be the discrimination of each probe $p \in P$ at hybridization $H_j \in H$, and let $\overline{R(p)} = \frac{1}{|H|} \cdot \sum_{H_j \in H} R_j(p)$ be the average discrimination of probe p over all hybridizations. We now compare $\overline{R(p)}$ of probe $p \in P$ to the free energy ΔG measured by $\mathcal{DP}(p,t)$ of probe–tag pair (p,t). Figure 12(a) depicts a scatterplot of $\overline{R(p)}$ versus ΔG for all probes of the HuGeneFL chip. As shown in this figure, we find that for probes p with matching tags t, $\overline{R(p)}$ varies between -0.8 and 1 if $\mathcal{DP}(p,t) \leq -22$. We find that $\overline{R(p)} \approx 0$ and varies between -0.219 and 0.099 if $\mathcal{DP}(p,t) > -22$. It must be the case that either (i) such probe–tag pairs have near zero average discrimination scores simply because the targets of these probes have not been differentially expressed over the > 120 experiments (in other words, there is a lack sufficient biological diversity), or (ii) such probe–tag pairs of HuGeneFL are too weak to discriminate between expressed and non-expressed states (and therefore should be removed from the chip or ignored). To rule out case (i) for the majority of such probe–tag pairs, we focus on the set of targets $G' \subseteq G$ such that if $g \in G'$, then there exists a distinct probe $p' \in P_g$ with $\mathcal{DP}(p',t') > -22$ for matching tag $t' \in T$. For $g \in G'$, if each probe $p \in P_g$ has $\overline{R(p)} \approx 0$, then we could conclude that the target g was not differentially expressed over the hybridization experiments. However, we find this is not the case since for the majority of targets $g \in G'$, at least one third of the probes $p \in P_g$ have $\overline{R(p)}$ much greater or much less than 0. Figure 12(b) depicts $\overline{R(p)}$ of all probes $p \in P$ such that $p \in P_g$ and $g \in G'$. The range of $\overline{R(p)}$ in Fig. 12(b) shows that many probes belonging to groups in G' exhibit average discriminations that are both much greater and much less than 0.

F12

As depicted in Fig. 13, we find that over all hybridization experiments, the discrimination $R_j(p)$ of all probes $p \in P$ with $\mathcal{DP}(p,t) > -22$ for matching tag t does not deviate greatly from $\overline{R(p)} \approx 0$, for all $H_j \in H$. Notice that the deviation of $R_j(p)$ from $\overline{R(p)}$ increases as $\mathcal{DP}(p,t)$ decreases, indicating that probes become more discriminatory as their hybridization strength increases. We conclude that the hybridization between a probe p and its matching tag t where $\mathcal{DP}(p,t) > -22$ is too weak for p to have a significant detection p-value (p is not able to discriminate).

F13

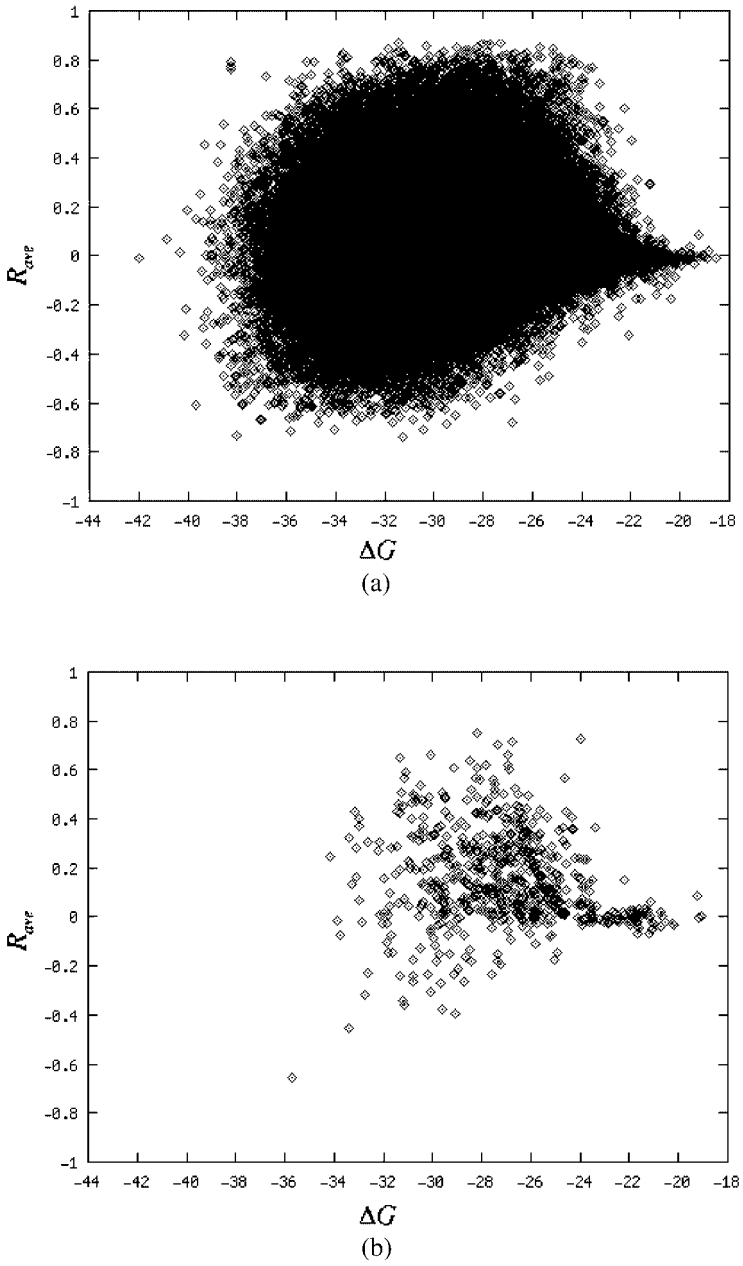


FIG. 12. (**a**) A scatterplot of $\overline{R(p)}$ versus $\mathcal{DP}(p,t)$ of all probes in $p \in P$ for HuGeneFL. (**b**) A scatterplot of $\overline{R(p)}$ against $\mathcal{DP}(p,t)$ of all probes $p \in P_g$ such that there exists a distinct probe $p' \in P_g$ with $\mathcal{DP}(p',t') > -22$, for matching tag t'.

7. OPEN PROBLEMS AND FUTURE DIRECTIONS

We present a framework for detecting degenerate probes in Affymetrix oligonucleotide microarrays. The predictions are based on a nearest neighbor model of hybridization. We show that the ΔG estimates from this theoretical model are strongly correlated with the distribution of ranks for a probe within its probe group over a large set of hybridization experiments. Each of four types of degenerate behavior induce four distinct distribution of ranks. The structural analysis of the conflict graph for the Affymetrix HuGeneFL chip produced several key insights that give us better prediction strategies. We find that very strong probe–tag pairs (low ΔG estimates) are more frequently predicted to be degenerate than are mid-range or weak

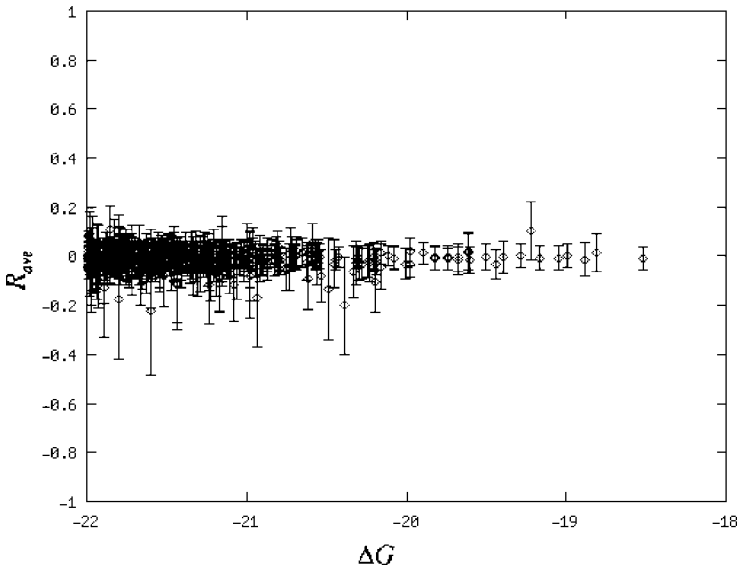


FIG. 13. This figure depicts $\overline{R(p)}$ and the standard deviation of $R(p)$ against $\mathcal{DP}(p,t)$ of all probes in $p \in P$ with $\mathcal{DP}(p,t) > -22$, for matching tag t.

probe–tag pairs (higher ΔG estimates). We see that the distribution of ranks of a probe is dependent on the hybridization strength between the probe and the matching tag.

We also give *support functions* for detecting whether a probe has an affinity for a particular degenerate behavior. These support functions do not make use of the theoretical models for hybridization but instead examine the distribution of ranks over a large set of hybridization experiments alone. These functions will better discriminate between degenerate and nondegenerate probes as the number and diversity of the hybridization experiments increases. A wide range of conditions guarantees a high degree of *biological diversity* (the expression of each target represented on the chip varies due to changes caused by the conditions under which the hybridization was performed). When high biological diversity is present for each probe in a set of hybridizations, there will be a large number of hybridizations in each block. This will lead to fewer zero values in elements of the rank count vectors. The experiments contained in this paper were carried out with a relatively small set of 126 hybridizations. A significantly larger (but manageably large) collection of hybridizations would ensure that sufficient biological diversity exists so that each probe is expressed (either highly or lowly) a substantial number of times.

An online resource containing additional results is publicly available (Smith and Hallett, 2004), and we will make our software freely available at this same location. Our software is sufficiently robust as to carry out these experiments with a ten-fold increase in the number of hybridization experiments.

Lastly, we show a strong correlation between the Affymetrix *discrimination* and ΔG estimates from the nearest neighbor model for a probe. In particular, we show that probes with high free energy (weak hybridization) have almost always a discrimination of 0. That is, their perfect match and mismatch probes intensities are the same, and hence, the probe is not informative.

Beyond simply increasing the number and diversity of hybridization experiments, it would of course be interesting to design better support functions that require fewer data. There is also the option of employing an alternative test statistic. In designing these experiments, we experimented with a chi-square test to determine the goodness of fit between the estimated background probability distribution and the observed data. The chi-square test gave no better results than the log likelihood ratio test statistic. We hope to refine our current test statistic and research other nonparametric methods for determining an underlying pattern from an observed rank count vector.

The focus of Smith and Hallett (2004b) is the integration of our framework with the model-based analysis of oligonucleotide arrays from Li and Wong (2001). Li and Wong give a simple model for determining the intensity measurement for a target as a nonlinear combination of the probe intensities from the probe group and parameters that specify the quality (sensitivity) of each probe. Both the theoretical model of

hybridization and the support functions based on the distribution of rank patterns presented in our paper can be modified to give a score for the quality of a probe. In this way, our framework provides an alternative, possibly better avenue for estimating parameters for use in the Li and Wong model.

This paper is primarily concerned with detecting patterns in hybridization experiments corresponding to degenerate probes. One can imagine several other useful patterns that would be of interest. For instance, a pattern could be designed for detecting correlation or causation in gene expression experiment data by examining the intensities of probe groups over the set of hybridizations. This approach could be used for network inference and finding network motifs (building blocks) of transcriptional regulation networks (Shen-Orr *et al.*, 2002).

Although we describe only an application of the model to the Affymetrix HuGeneFL chip, the model has been designed to be universal, so that it can be used to analyze the quality of any existing oligonucleotide microarray or microarray design. We are currently comparing several different Affymetrix GeneChips (Smith and Hallett, 2004b) in order to determine whether there is quantitative evidence that these chips are gradually minimizing the amount of degenerate behavior. Ultimately, to validate the model, the candidate degenerate probes and probes pairs must be verified in a wet-lab to conclude whether they are truly degenerate. We are currently designing a chip containing a wide variety of degenerate and nondegenerate probes and probe pairs in such a way that very few hybridization experiments will be necessary to validate the model.

REFERENCES

Affymetrix (TM). 2002. White Paper: *Statistical Algorithms Description Document*, Affymetrix, Inc.

BenDor, A., Karp, R., Schwikowski, B., and Yakhini, Z. 2000. Universal DNA tag systems: A combinatorial design scheme. *J. Comp. Biol.* 7(3/4), 503–519.

Hubbell, E., and Pevzner, P. 1999. Fidelity probes for DNA arrays. *Proc. 7th Int. Conf. on Intelligent Systems for Molecular Biology*, 113–117.

Lemon, W.J., Palatini, J.J.T., Krahe, R., and Wright, F.A. 2002. Theoretical and experimental comparison of gene expression indexes for oligonucleotide arrays. *Bioinformatics* 18(11), 1470–1476.

Li, C., and Wong, W. 2001a. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci*. 98(1), 31–36.

Li, C., and Wong, W. 2001b. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol*. 2(8), 1–11.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* 288, 910–940.

Moore, D.S. 1995. *The Basic Practice of Statistics*, W.H. Freeman, New York.

NetAffx (TM) website. 2002. *www.netaffx.com*.

Sankoff, D., and Zuker, M. 1984. RNA Secondary structures and their prediction. *Bull. Math. Biol.* 46, 591–621.

SantaLucia, Jr., J. 1998a. A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95, 1460–1465.

SantaLucia, Jr., J. 1998b. Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with Watson–Crick base pairing. *Amer. Chem. Soc. Biochem.* 37, 14719–14735.

SantaLucia, Jr., J. 2002. Loop parameters. Unpublished data. *www.bioinfo.math.rpi.edu/žukerm/dna/credit.html*.

SantaLucia, Jr., J., Allawi, H., and Seneviratne, A. 1996. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Amer. Chem. Soc. Biochem.* 96(35), 3555–3562.

Sengupta, R., and Tompa, M. 2000. *Quality control in manufacturing oligo arrays: A combinatorial design approach*. Technical Report #2000-08-03, Department of Computer Science and Engineering, University of Washington.

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* 31, 64–68.

Ship, M., Ross, K., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., and Golub, T.R. 2002. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine* 8(1).

Smith, K., and Hallett, M. 2004a. Online resource for paper. *www.mcb.mcgill.ca/~genechips*, McGill University.

Smith, K., and Hallett, M. 2004b. Estimating probe parameters for intensity computations. In preparation.

Sugimoto, N., Nakano, S., Yoneymana, M., and Honda, K. 1996. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucl. Acids Res.* 24(22), 4501–4505.

Tobler, J.B., Molla, M.N., Nuwaysir, E.F., Green, R.D., and Shavlik, J.W. 2002. Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *Intelligent Systems for Molecular Biology (ISMB 2002)*, 164–171.

Virtaneva, K., Wright, F.A., Tanner, S.A., Yuan, B., Lemon, W.J., Caligiuri, M.A., Bloomfield, C.D., de la Chapelle, A., and Krahe, R. 2001. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl. Acad. Sci.* 98(3), 1124–1129.

Zuker, M., Mathews, D.H., and Turner, D.H. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. *In* J. Barciszewski and B.F.C. Clark, eds., *RNA Biochemistry and Biotechnology*, NATO ASI Series, Kluwer Academic Publishers, Amsterdam.

Address correspondence to:
Mike Hallett
McGill Centre for Bioinformatics
Duff Medical Building
3775 University Street
McGill University
Montreal, Canada

E-mail: hallett@mcb.mcgill.ca

AU1
Change okay: “is that probe”?

AU2
Change okay: “that there is a higher number”?

QU3
Okay to number level 3 headings in Section 5.1 per style of other sections?