Towards Quality Control in DNA Microarrays

Kaleigh Smith
McGill Centre for Bioinformatics
School of Computer Science
McGill University, Montreal, Canada
`kaleigh@mcb.mcgill.ca`

May 6, 2003

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Masters of Science.

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-88298-5
Our file Notre référence
ISBN: 0-612-88298-5

Canada

Abstract

We present a framework for detecting degenerate probes in a DNA microarray that may add to measurement error in hybridization experiments. We consider four types of behaviour: secondary structure formation, self-dimerization, cross-hybridization and dimerization. The framework uses a well-established model 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 correlated with each type of degenerate probe behaviour. The framework also contains a machine learning technique to learn from the hybridization experiment data. We implement the components of the framework and evaluate the ability of the framework to detect degenerate probes in the Affymetrix HuGeneFL GeneChip.

Résumé

La présence de sondes dégénerées dans des puces à ADN contribue à l'erreur des mesures lors d'expériences d'hybridation. Nous présentons un système qui permettra la détection de sondes dégénerées exhibant un des quatre types de comportements non-désirables: la formation de structures secondaires, l'auto-dimérisation, la trans-hybridation et la dimérisation. Notre systéme emploie un modèle d'hybridation d'acides nucléiques bien connu ainsi qu'une nouvelle méthode de détection de motifs répétés dans des données d'expériences d'hybridation. Notre contribution principale à l'avancement des recherches dans ce domaine est la description de motifs d'hybridation spécifiques à chacun des quatre types de comportements dégénerées des sondes que nous étudions. Notre systëme utitlise une technique d'intelligence artificielle qui permet l'apprentissage

à partir des données des experiences d'hybridation. Nous avons évalué la capacité de détection des sondes dégénerées de notre système en analysant la puce HuGeneFL GeneChip d'Affymetrix.

Acknowledgement

This thesis could not have been completed without the imagination, guidance, collaboration and support of my advisor Dr. Michael Hallett. I thank him for his vision of this project and his confidence that I could succeed. The professors and students at the McGill Centre for Bioinformatics were a constant source of energy and encouragement, particularly my good and bright friend Louigi Addario-Berry. Feedback and questions from students, namely Francois Pepin and Alex Marcil, helped to shape and develop this research. Thanks to Michelle Scott for translating the abstract for this thesis into French. To Eragen Biosciences, for their code and introduction to the problems of DNA Microarray chip design. I'd also like to thank the researchers at Affymetrix and various universities who shared their knowledge and microarray data online. Personally, I've been surrounded by an ardent team of supporters and friends, my parents Bill and Lynn, sister Brennagh, the Mikula's and Krzystyna Batko. And for their love and humour during 2002, Alex Batko and Amy Goalen.

Contents

List of Tables

List of Figures

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 present in the measurements of gene expression produced by this technology represents a significant obstacle to understanding gene regulatory mechanisms. This thesis 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 a short strand of nucleic acids (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.[1] In an experiment, the mRNA is preprocessed and labelled 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 (termined *probe-tag pair*) is optically measured. In an error-free scenario, this intensity is proportional to the

---

[1] Appendix A contains a brief description of hybridization.

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.

Variability among different hybridization experiments is commonly characterized as biological variability, sample variability or technical variability. We are concerned primarily 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. Bias error will be consistent over all hybridizations. We focus on the development of tools for detecting and eventually predicting bias error in experiments.

Several strategies exist in the literature for designing DNA microarrays that attempt to minimize (or detect) various types of error [2, 6, 18, 26]. 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 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').

Consider secondary structure degeneracy where a single strand of nucleic acid has the ability to fold and hybridize into a stable folded secondary structure. As we focus on short sequences, we consider only the most simple folded structure called a *hairpin loop.* For example, tag $t = AAAAAAAAGGGTTTTTTTT$ could fold to form a stable hairpin structure as depicted in Figure 1. Now consider tags that exhibit self-dimerization. For instance, when two copies of $t = TTTCCCATGGGAAA$ are present in a solution, they may hybridize to form a stable duplex as depicted in Figure 2 Tags may hybridize to unintended probes causing cross-hybridization. Consider $t = TTACCCTTAAGTAC$, $t' = TTACCGTTAAGTAC$ and $p = GTACTTAAGGGTAA$. t' may hybridize to p as depicted in Figure 3 (note also that t may hybridize to p'). Finally, different tags $t = TTACCGTTAAGTAC$ and $t' = GTATTTAACGATAA$ may dimerize as depicted in Figure 4.

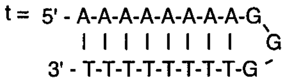


Figure 1: An example of a hairpin loop.

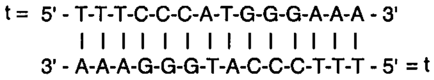

t = 5' - T-T-T-C-C-C-A-T-G-G-G-A-A-A - 3'
3' - A-A-A-G-G-G-T-A-C-C-C-T-T-T - 5' = t

Figure 2: An example of self-dimerization.

A well designed microarray should prevent the occurrence of these four types of degeneracy. Figure 5 depicts a stylized view of the four types of degeneracies that may

```
t' = 5' - T-T-A-C-C-G-T-T-A-A-G-T-A-C - 3'
         | | | | |   | | | | | | | | |
       3' - A-A-T-G-G-G-A-A-T-T-C-A-T-G - 5' = p
```

Figure 3: An example of cross-hybridization.

```
t =  5' - T-T-A-C-C-G-T-T-A-A-G-T-A-C - 3'
         | | |   | | | | | | |   | | |
       3' - A-A-T-A-G-C-A-A-T-T-T-A-T-G - 5'  = t'
```

Figure 4: An example of dimerization.

occur during a hybridization experiment. Note that we do not claim that all of the expressed tags will hybridize to the probes. In hybridization experiments, it has been observed that only a proportion of the expressed tags will hybridize to the probes on the chip [4]. This observation does not affect the work described here, as we examine situations when fewer than the expected proportion of tags hybridize to a probe on the chip.

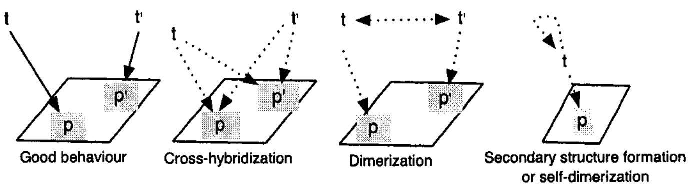


Figure 5: Solid arrows indicate that the expected amount of expressed tags should participate in the hybridization. Dotted arrows indicate that some fraction of the expected amount of available tags may participate in the hybridization.

Each of the four degenerate behaviours may contribute to error in gene expression measurements in distinct ways. As a simple example, consider the effect of secondary structure on the intensity of a probe p_1. Consider a set of tags $t_1, \ldots, t_l$ representing some target g where t_1 is known 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. Similar style arguments exist for the remaining types of degeneracy.

We conjecture that each of the four types of degeneracy will create unique patterns in the hybridization data. The unique pattern is recognizable by comparing the intensity of a probe to the intensities of the remaining probes in its group. For example, a plausible pattern of degenerate behavior caused by secondary structure is illustrated as follows. Considering 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 $t_2, \ldots, t_l$, then it is possible that t_1 is prone to form secondary structure and will not hybridize with p_1 regardless of the condition under which the hybridization is performed. If the rank of the intensity of this probe is consistently low w.r.t. probes $p_2, \ldots, p_l$ especially when target g is highly expressed, then this probe-tag pair exhibits a pattern attributed to secondary structure. By using appropriate patterns for secondary structure, cross-hybridization, dimerization and self-dimerization, we show evidence that specific probe-tag pairs do behave according to these patterns.

In Section 2, we show how to build a *conflict graph* for a chip. This conflict graph helps us to organize the relationships between the probes. Each vertex in the conflict graph corresponds to a probe on the chip. A self-loop (p, p) is added to the graph if the tag t associated with probe p exhibits a high affinity to form strong secondary structure or to self-dimerize. An edge (p, p') is added between distinct probes, if the associated tags t and t' exhibit a high affinity to cross-hybridize with p' and p or if tags t and t' exhibit a high affinity to dimerize. The affinity for a probe (or pair of probes) to exhibit each type of degeneracy is estimated by 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) [10, 13, 15, 17, 24, 28]. For dimerization and cross-hybridization, we use a straightforward dynamic programming algorithm to calculate the minimum free-energy over all possible hybridizations between the two sequences with the *nearest-neighbor* model and thermodynamic parameters supplied by [15]. In all cases, an edge is added to the conflict graph if and only if the minimum free-energy exceeds a conservative threshold.

In Section 3 we provide a method and notation for formally defining patterns in hybridization experiments. We begin by examining the pattern for probes not affected by any type of degeneracy (*non-degenerate* probes). We then define the unique patterns associated with each of the four degenerate behaviours and the intuition behind each pattern. We finish by justifying the patterns by showing that probes measured to have high affinity to be degenerate do in fact exhibit such patterns.

In Section 4 we show how the definitions of the unique patterns of degeneracy and non-degeneracy can be used to measure the extent to which hybridization data supports a claim that a particular probe (or pair of probes) is degenerate.

Section 5 extends the conflict graph to allow for learning from hybridization data. As the chip is increasingly used by the community and the raw intensity values (i.e., the cell files) are made publically available, the potential for learning this library of knowledge is increased. We describe a learning approach to detect degenerate probes or probe pairs. The method seeks to identify patterns in a training set of hybridization experiments that may be indicative of degenerate probes. We discuss methods for learning from hybridization data and the challenges that ensue.

Throughout this work, we test our framework with the *Affymetrix HuGeneFL* chip [12] which contains on the order of 10^5 probes of length 25 bp representing approximately $6,000$ human targets. This is a relatively old chip and the data from many hybridization experiments is publically available. The initial experiments described here use a set of 126 hybridizations from three laboratories [7, 20, 27]. We also work within the framework to examine the structure of the conflict graph and to investigate additional properties of the HuGeneFL chip such as the discrimination level of probes.

2 Conflict Graph Framework

The following notation is used for describing probes and their organization on a specific chip. Let $\Sigma_{dna} = \{A, C, G, T\}$ and $\Sigma_{rna} = \{A, C, G, U\}$. A *probe* p of length n is a string $p = p_1p_2 \dots p_n \in \Sigma^n_{dna}$. A *tag* t of length n is a string $t \in \Sigma^n_{rna}$. The *reverse* t^r of t is the string $t_nt_{n-1} \dots t_1$. 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, $t_i \in \Sigma^n_{rna}$. 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 signficance). 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. 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 composed 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 $g \in G$.*

2.1 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 a an edge labelled 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 labelling the edges of M, $\kappa : E \longrightarrow \{s, x, d\}$, and $\tau = \{\tau_s, \tau_x, \tau_d\}$ is a set of suitable threshold parameters for the $\mathcal{S}$, $\mathcal{X}$, $\mathcal{D}$ functions.

Functions $\mathcal{S}$, $\mathcal{X}$ and $\mathcal{D}$ estimate the affinity for a probe or pair of probes to form a structure or duplex that may cause one of the four degenerate behaviours. Affinity for duplex formation is most commonly measured in terms of duplex stability or hybridization strength by *free-energy* ΔG [15] defined as the total change in energy from duplex to single-stranded states. Appendix B contains information on the measurement of duplex stability. In a series of papers, SantaLucia et. al. [14, 15, 16, 17] determine the thermodynamic hybridization parameters for most DNA 2-mers against most 2-mers including both wc-complentary 2-mers and mismatch 2-mers, and various misalignments. 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.

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 $5 - 60$ [14].

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

The algorithm denoted by $\mathcal{DP}$ uses standard dynammic programming with a constant gap penalty. As a bulge betwen 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 vs. RNA alignment whereas the computation of free-energy between a probe $p \in P$ and a tag $t \in T$ is a DNA vs. RNA alignment. Since the complete parameters, including mismatch parameters, required to compute the free-energy for RNA vs. DNA and RNA vs. RNA alignments were not publicly available, we use the parameters for DNA vs. DNA alignments as a good approximation [17]. Therefore, sequences input to $\mathcal{DP}$ may be over Σ_{dna} or Σ_{rna}, but in the latter case, the sequence will be translated 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 the actual cRNA tag. 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 a quantitative 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 [13] 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, the program returns the minimum free-energy over all hairpin structures (without pseudo-knots) for a nucleic acid sequence. We represent this as function $\mathcal{S} : \Sigma^n_{dna} \to \mathbb{R}$. $\mathcal{S}(p)$ is the affinity for probe p to fold into a secondary structure. Although $\mathcal{S}(p)$ only measures the affinity of p to form secondary structure, it is used to indicate that probe p is degenerate due to secondary structure formation of either p or the matching tag of p, t. In reality, the affinity for t to form secondary structure will differ from the affinity for p to form secondary structure for several reasons. For example, probe p is tethered to the chip while tag t is not. Furthermore, tag t is an RNA sequence and the free-energy of a secondary structure of t will be affected by a tailing ribonucleic acid sequence. However for simplicity, we assume that $\mathcal{S}(p) \approx \mathcal{S}(t)$ and we will write that a probe p has high affinity to form secondary structure to mean either t or p may form secondary structure. More formally, we include an edge $(p, p) \in E$, $p \in P$, if $\mathcal{S}(t) < \tau_s$, where t matches p and set $\kappa(p, p) \leftarrow s$. We determine the ΔG thresholds, including τ_s, below.

We use the algorithm $\mathcal{DP}$ to predict the pairwise behaviour 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 $\mathcal{X}(p,p') = min(\mathcal{DP}(p,t'), \mathcal{DP}(p',t))$ computes the affinity for t' to hybridize with p and t to hybridze with p'. If $\mathcal{X}(p,p') < \tau_x$, then we include an edge $(p,p') \in E$ and set $\kappa(p,p') \leftarrow x$. Similarily, let $\mathcal{D}(p,p')$ be the result of computing $\mathcal{DP}(t,t')$ where t, t' matches p, p' respectively. 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 dimerization, we allow that $p = p'$. In this case of *self-dimerization* (two copies of t dimerize), we add a self-loop to p and assign $\kappa(p,p) \leftarrow d$. For ease of notation, we say that probe p has an affinity to self-dimerize in such a case. We may also wish to introduce a threshold for self-dimerization τ_{sd} that differs from τ_d. These sets of secondary structure, self-dimerization, cross-hybridization, and dimerization edges are denoted by S, SD, X, and D respectively.

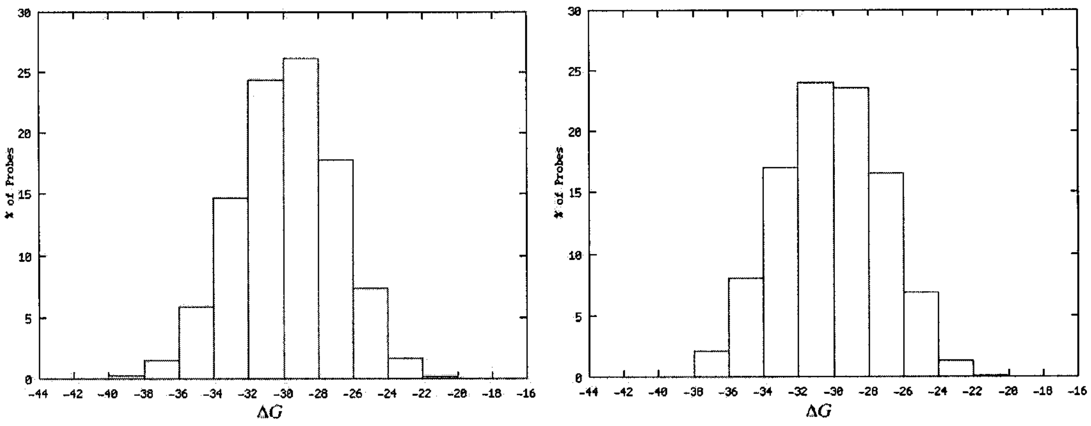
2.2 Conflict Graph for the Affymetrix HuGeneFL Chip

Our experiments are based on a set of 126 Affymetrix Inc. (TM) hybridization experiments made publically available by [7, 20, 27]. These hybridization experiments used the Affymetrix HuGeneFL chip (TM) and we use the information on the construction of HuGeneFL made available through NetAffx (TM) [12]. HuGeneFL is designed for human and has a probe set P' of size 131542 used to represent a set G' of 7129 genomic targets or groups. Affymetrix defines a probe group P_g for each $g \in G'$. For our purposes, we let C represent HuGeneFL by letting $\{G \subset G' : \forall g \in G, |P_g| = 20\}$ and $\{P \subset P' : \forall p \in P, p \in P_g,\ g \in G\}$. Tag set T and tag groups $\{T_g : g \in G\}$ are derived from P and $\{P_g : g \in G\}$. In this representation, $|G| = 6378$ and $|P| = 127560$ of which 127386 are unique DNA sequences. G mostly contains groups that target genes, and not groups used as controls or repeat detectors. However, 58 groups in G are in fact Affymetrix control groups. We include these in our experiments since they contain probes that may exhibit degenerate behaviour.

2.2.1 Analysis of ΔG

Consider first the function $\mathcal{DP}$ for calculating free-energy (ΔG). We calculate ΔG for 10^8 randomly chosen probes $p \in \Sigma_{dna}^{25}$ and their wc-complements $\bar{p}$ and find that -45.083 kcal/mol $\leq \mathcal{DP}(p, \bar{p}) \leq -15.603$ kcal/mol for all p. The range and distribution of ΔG of the random set of probes should, with high probability, represent the range and distribution of ΔG of all length 25 probes and their wc-complements. We use the ΔG information of the random set of probes as a standard against which to compare the ΔG of probe-tag pairs of chip C. We find that the distribution of ΔG of probe-tag pairs of C fits the distribution of all possible ΔG. The range of ΔG is -42.003 kcal/mol $\leq \mathcal{DP}(p, t) \leq -18.993$ kcal/mol for all probe-tag pairs (p, t), $p \in P$ and $t \in T$ of chip C. Figure 6 depicts histrograms of the ΔG of these two sets of probes.

Let τ_h be the floor of the largest ΔG over all probe-tag pairs (p, t), $p \in P$ and $t \in T$. Here, $\tau_h = -19$. Note that threshold τ_h represents the ΔG of the weakest probe-tag pair of C. We assume that the ΔG of any hybridization that occurs between a probe and a tag in $P \times T$ or a pair of tags in $T \times T$ is at most τ_h. We calculate cross-hybridization ΔG (measured by function $\mathcal{X}$) and dimerization ΔG (measured by function $\mathcal{D}$) of all pairs of probes $(p, p') \in P \times P$, $p \neq p'$ to find pairs of probes such that $\mathcal{X}(p, p') \leq \tau_h$ or $\mathcal{D}(p, p') \leq \tau_h$. Of the more than 8×10^9 possible pairs of probes, 239 pairs have $\mathcal{X} \leq \tau_h$ and 487 pairs have $\mathcal{D} \leq \tau_h$. Figure 7 depicts two histograms of ΔG values measured by $\mathcal{X}$ at most τ_h and ΔG values measured by $\mathcal{D}$ at most τ_h. The follow-

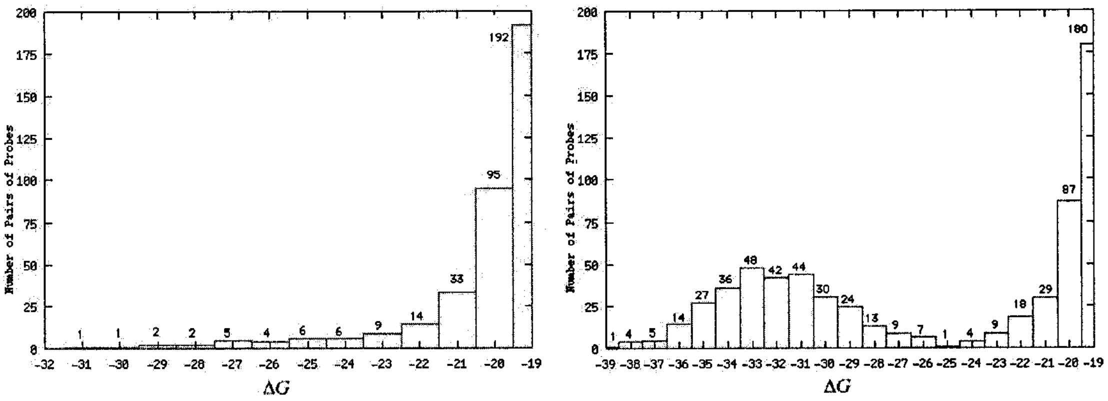


(a)

(b)

Figure 6: Histrogram (a) depicts the percentage of ΔG values measured by $\mathcal{DP}(p, \bar{p})$ for 10^8 randomly chosen probes $p \in \Sigma^{25}_{dna}$, and histogram (b) depicts the percentage of ΔG values measured by $\mathcal{DP}(p, t)$ for all probe-tag pairs (p, t), $p \in P$ and $t \in T$ of chip C.

ing are examples of pairs of probes measured to have low $\mathcal{X}(p, p')$ or low $\mathcal{D}(p, p')$. 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 one base and belong to different probe groups. As an example of possible dimerization, probe $p = CGAAGCGGAATTCTCCATGCCCGAG$ in probe group M24899_at and probe $p' = CTCGGGCATGGAGAATTCCGCTTCG$ in probe group X72632_s_at are measured to have $\mathcal{D}(p, p') = -32.4935$. Note that probes p and p' are wc-complements. We find many other similar examples and conclude that $\mathcal{X}$ and $\mathcal{D}$ are capable of detecting pairs of probes with an affinity for cross-hybridization or dimerization.

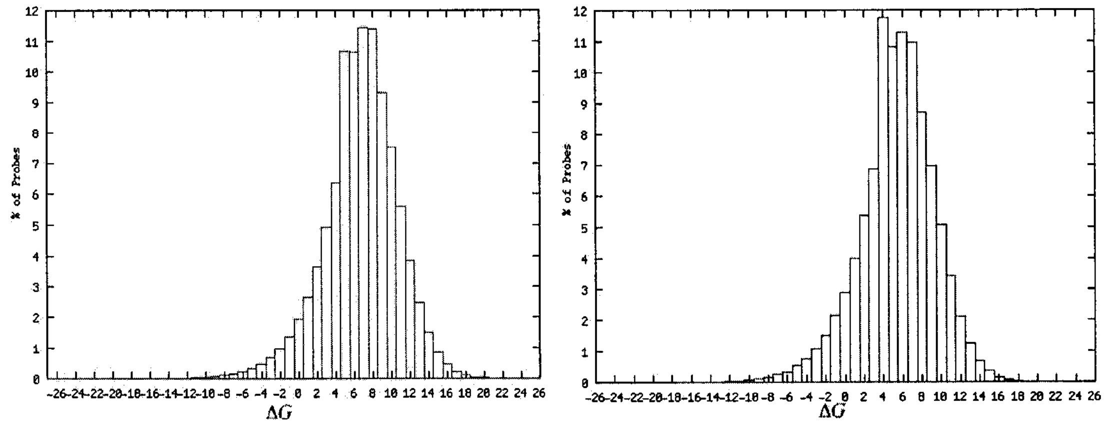


(a) (b)

Figure 7: Histogram (a) depicts the number of ΔG values at most τ_h measured by $\mathcal{X}(p,p')$ for $p,p' \in P, p \neq p'$, and histogram (b) depicts the number of ΔG values at most τ_h measured by $\mathcal{D}(p,p')$ for $p,p' \in P, p \neq p'$.

Consider now the special case of dimerization between two copies of the same probe (self-dimerization). We would first like to know the range of self-dimerization ΔG over all probes of length 25. We remove the restriction that $\mathcal{D}(\cdot,\cdot) \leq \tau_h$ and calculate $\mathcal{D}(p,p)$ for 10^8 randomly chosen probes $p \in \Sigma_{dna}^{25}$. We also calculate $\mathcal{D}(p',p')$ for all probes $p' \in P$. Figure 8 depicts two histograms of $\mathcal{D}(\cdot,\cdot)$ over both sets of probes. We find that $-16.303 \leq \mathcal{D}(p',p') \leq 22.286$ for all probes $p' \in P$, compared to $-27.763 \leq \mathcal{D}(p,p) \leq 25.716$ for 10^8 randomly chosen probes $p \in \Sigma_{dna}^{25}$. Since the two distributions in Figure 8 are similar, there exist probes predicted by $\mathcal{D}$ to have a high affinity for self-dimerization (those with low ΔG). However, the lowest ΔG for self-dimerization over all probes $p \in P$ is much higher than the lowest possible ΔG values for self-dimerization. Note that we find that there is no probe $p \in P$ such that $\mathcal{D}(p,p) \leq \tau_h$. An example of a probe measured to have low $\mathcal{D}$ 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 may hybridize with only four mismatches.



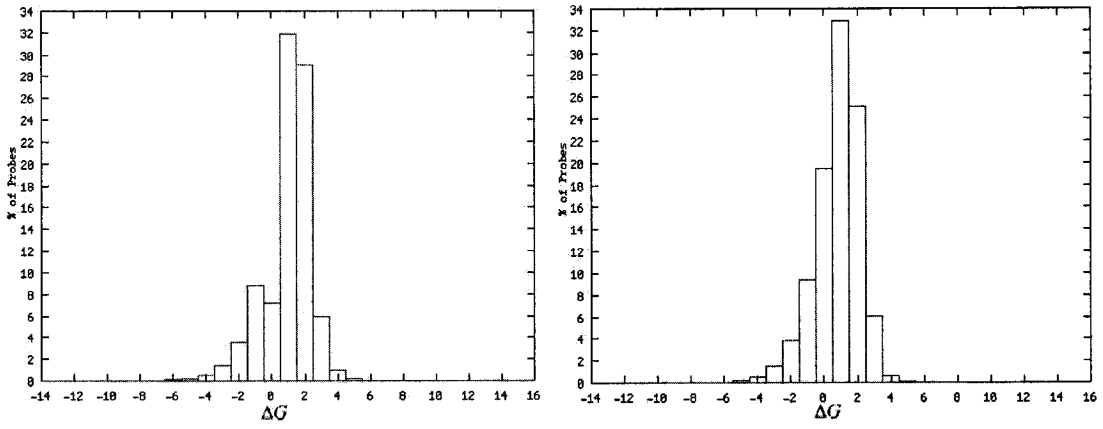
(a)

(b)

Figure 8: Histogram (a) depicts the percentage of ΔG values measured by $\mathcal{D}(p,p)$ for 10^8 randomly chosen $p \in \Sigma^{25}_{dna}$, and (b) depicts the percentage of ΔG values measured by $\mathcal{D}(p,p)$ for all probes $p \in P$.

The range of ΔG for self-hybridization (secondary structure formation) of a probe is different than that of hybridization between a probe and its wc-complement. This motivates us to calculate $\mathcal{S}(p)$ for 10^8 randomly chosen probes $p \in \Sigma^{25}_{dna}$ to serve as a standard against which to compare $\mathcal{S}(p')$ for probes $p' \in P$ from chip C. As with self-dimerization, the lowest ΔG measured with $\mathcal{S}$ over set P is much higher that the lowest ΔG measured by $\mathcal{S}$ over the random set of probes. We find that $-8.678 \leq \mathcal{S}(p') \leq 8.525$ for all probes $p' \in P$, compared to $-14.579 \leq \mathcal{S}(p) \leq 11.335$ for 10^8 randomly chosen probes $p \in \Sigma^{25}_{dna}$. Figure 9 depicts two histograms of $\mathcal{S}$ over a random set of probes and over all probes in P. Examples of probes measured to have low $\mathcal{S}$ are $p = GCCACCACACTGGTGTGCTGGCTGT$ with $\mathcal{S}(p) = -8.67883$ and

$p' = GCGAGGAAGCTTCCTCGCAACTTTG$ with $\mathcal{S}(p') = -7.36687$. Both p and p' can fold in such a way (not necessarily at the middle of the sequence) to form a hybridization with few mismatched base pairs.



(a)

(b)

Figure 9: Histrogram (a) depicts the percentage of ΔG values calculated by $\mathcal{S}(p)$ for 10^8 randomly chosen $p \in \Sigma_{dna}^{25}$, and histogram (b) depicts the percentage of ΔG values calculated by $\mathcal{S}(p)$ for all probes $p \in P$.

2.2.2 Analysis of Conflict Graph

The edge labelled conflict graph M is created from chip C and edges are added by applying functions $\mathcal{S}, \mathcal{X}$ and $\mathcal{D}$ to probes in P as detailed in Section 2.1. ΔG thresholds for each function (τ_s, τ_x and τ_d) are selected according to the ranges of ΔG over a set of random length 25 probes and over set P. We include a new threshold for self-dimerization, τ_{sd}, since we needed a non-empty set SD to study probes with an affinity for self-dimerization. The sizes of edge sets S, SD, X and D for a variety of such thresholds are depicted in Table 1. Based on the sizes of edge sets S, SD, X and D and the range of ΔG over all probe-tag pairs of chip C, we create a conflict graph M with $\tau_s = -6$, $\tau_{sd} = -14$, $\tau_x = -23$ and $\tau_d = -33$.

τ_s (kcal/mol)	$\|S\|$	τ_{sd} (kcal/mol)	$\|SD\|$	τ_x (kcal/mol)	$\|X\|$	τ_d (kcal/mol)	$\|D\|$
-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 1: Edge set sizes of HuGeneFL conflict graph M compared with various threshold values in kcal/mol.

Conflict graph M contains no vertices with more than one self-loop. Therefore no probes are predicted to have an affinity for both secondary structure formation and self-dimerization. Let $S \subseteq P$ and $SD \subseteq P$ be the probes corresponding to vertices with self-loops labelled s and labelled d respectively. The average ΔG measured by $\mathcal{DP}$ of probe-tag pairs (p, t) is -33.67294 for $p \in S$ and -32.52438 for $p \in SD$. Therefore, probes in S and SD typically have matching tags such that the probe-tag pair has low ΔG (strong probe-tag pairs). We now consider vertices adjacent to edges indicating an affinity for cross-hybridization or dimerization. Let M_x be M with edge set E restricted to edges labeled x. It is the case that the majority of non zero-degree vertices in M_x have degree one, although the maximum degree of M_x is three. Let M_d be M with edge set E restricted to edges labeled d and excluding self-loops. For the HuGeneFL chip, all the vertices in M_d have either degree zero or one and there are no vertices in $M_{xd} = M_x \cup M_d$ incident to both x and d labelled edges.

Vertex Degree	M_x	ave. ΔG	M_d	ave. ΔG	M_{xd}	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

Table 2: Number of vertices with degree i and average probe-tag pair strengths ave. ΔG for probes represented by vertices with degree i of HuGeneFL conflict graph $M = (V, E)$ for subgraphs M_x, M_d and M_{xd}.

3 Patterns in Hybridization Experiments

The idea that degeneracy will create unique patterns in the hybridization data was introduced in Section 1. Informally, a pattern was described by the intensity of a probe compared to the intensities of the remaining probes in its group. The following notation and definitions formalize the concept of patterns in hybridization data.

3.1 Notation and Definitions

Let $\mathcal{H} = \{H_1, \ldots, H_K\}$ be the set of hybridization experiments. The output of a hybridization experiment consists of an intensity value for every probe on the chip. 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, where G and P_g are defined by the chip C. The *intensity of target* $g \in G$ *in experiment* H_j is represented by $I_j(g) \in \mathbb{R}$. In a standard hybridization experiment, the expression level of a target is determined via a standard statistical method taking into account, for instance, specificity of the probe-tag pair. We require 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 training set of hybridizations, the intensity measurements of a probe (of a target) for all hybridizations are scaled to the $(0 \ldots 1]$ interval. More robust variants of $I_j(g)$ are possible. For example, we could normalize data beforehand or remove certain probes from the analysis. We do not discuss such strategies any further in this thesis.

At each experiment $H_i \in \mathcal{H}$ and each target $g \in G$, the intensity of each probe

$p \in P_g$ is *ranked* relative to the intensity of all remaning 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* $p_1, \ldots, p_{i-1} \in P_g \backslash \{p\}$ *s.t.* $I_j(p_k) < I_j(p)$, *for* $1 \leq k \leq i-1$. *When the probe group is clear from the context, we denote the rank simply as* $\rho_j(p)$.

We wish to discretize the hybridization experiments into blocks according to $I_j(g)$. Let $b \in \mathbb{Z}$ be 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* $\frac{b'-1}{b} < I_j(g) \leq \frac{b'}{b}$.

The following definitions relate the rank of a probe p to the intensity of the target of p. Throughout the remainder of this thesis, we assume 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* $p \in P_g$, $1 \leq i \leq l$ *and* $1 \leq j \leq K$.

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

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

Definition 6 (Rank Count Vector) *For* $P' \subseteq P$ *and* $H \subseteq \mathcal{H}$ *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* $H \subseteq \mathcal{H}$ *, 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) is a discrete description of the pattern made by the ranks of a set of probes over a set of hybridization experiments. However we are primarily interested in the underlying pattern (distribution) of ranks than the number of occurrences of each rank. We require a family of distributions. Some members of the family describe a pattern where the majority of ranks are high and very few ranks are low (exponential distribution tailing to the left), a symmetric pattern where the majority of ranks are low and very few ranks are high (exponential distribution tailing to the right), a pattern of equal ranks (uniform distribution) and a variety of other patterns. The *beta distribution* with parameters α and β is appropriate for this purpose 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 [5].

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

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

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

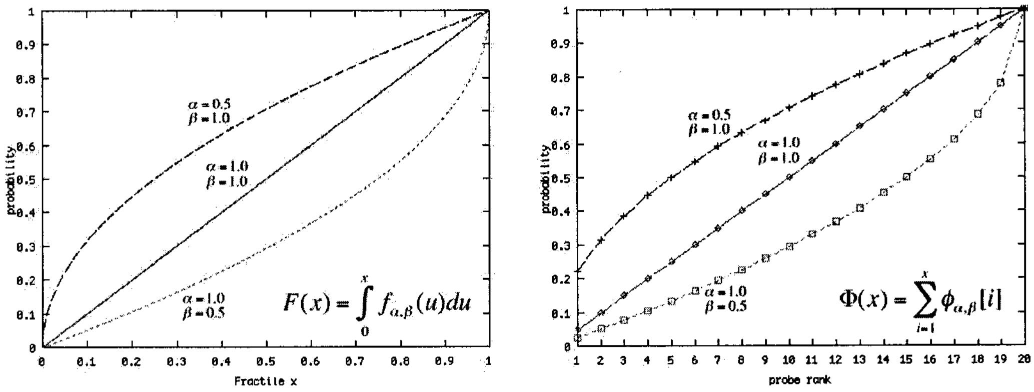
We would also like to have the ability to compare how will a rank count vector fits a particular beta distribution. Towards this end, we require a discretization of the continuous beta distribution.

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*

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

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

Figure 10 depicts the distribution function of beta distributions for a variety of parameters α and β and the distribution function of the discretized probability vectors derived from the beta distributions as shown in Definition 8. Example 1 is a small example of the algorithm for determining a discretized probability vector from rank count vectors for a set of probes.



(a) (b)

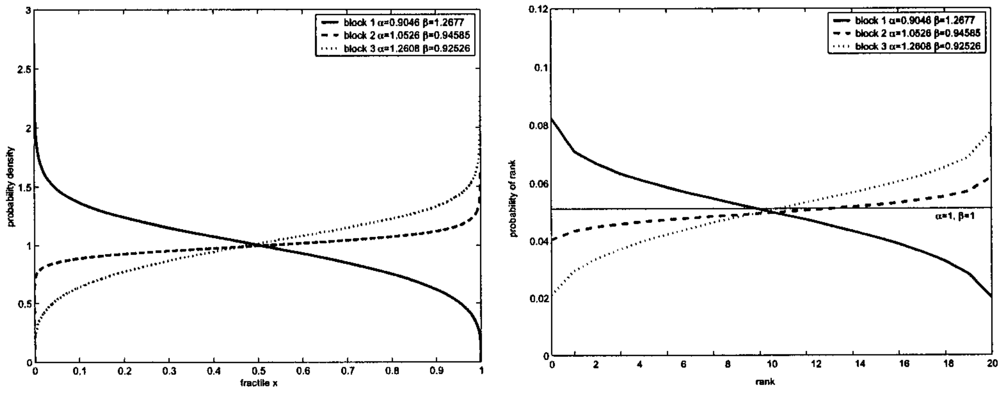
Figure 10: Distribution functions for three different values of α and β where (a) $F(x)$ is the distribution function of the beta density function $f_{\alpha,\beta}(u)$ and (b) $\Phi(x)$ is the distribution function discretized probability vector $\phi_{\alpha,\beta}$.

Example 1 *Let $P' \subseteq P$ be some set of probes and $H \subseteq \mathcal{H}$ be a set of hybridization experiments. Let $b = 3$ be the number of blocks. The beta distribution α and β parameters are estimated from three rank count vectors $y_1^{P',H}$, $y_2^{P',H}$ and $y_3^{P',H}$. Consider the following three rank count vectors:*

$$
\begin{aligned}
y_1^{P',H} &= \langle 3882, 3542, 3047, 2902, 2624, 2420, 2212, 2168, 2029, 1970, 1934, 1809,\\
&\quad 1620, 1399, 1439, 1348, 1715, 1391, 1052, 533\rangle\\
y_2^{P',H} &= \langle 685, 684, 685, 700, 640, 685, 685, 685, 685, 685, 685, 685, 685, 685, 685, 741\\
&\quad 685, 685, 685, 685\rangle\\
y_3^{P',H} &= \langle 533, 1052, 1391, 1715, 1348, 1439, 1399, 1620, 1809, 1934, 1370, 2029, 2168,\\
&\quad 2212, 2420, 2624, 2902, 3047, 3542, 3882\rangle.
\end{aligned}
$$

The α and β parameters estimated from these samples according to Equation 1 are $\widehat{\alpha}_1 = 0.9046$, $\widehat{\beta}_1 = 1.2677$, $\hat{\alpha}_2 = 1.0526$, $\widehat{\beta}_2 = 0.94585$, $\hat{\alpha}_3 = 1.2608$ and $\hat{\beta}_3 = 0.92526$. Figure 11 depicts the three discretized probability vectors $\phi_{\alpha_{b'},\beta_{b'}}$ for estimated $\widehat{\alpha}_{b'}$ and

$\widehat{\beta}_{b'}$, $1 \leq b' \leq 3$.



(a)

(b)

Figure 11: Parameters $\widehat{\alpha}_{b'}$ and $\widehat{\beta}_{b'}$ estimated from the rank count vectors of Example 1, for $1 \leq b' \leq 3$. Figure (a) depicts the three beta distributions curves $f_{\widehat{\alpha_{b'}},\widehat{\beta_{b'}}}$, and figure (b) depicts the three discretized probability vectors $\phi_{\widehat{\alpha_{b'}},\widehat{\beta_{b'}}}$.

3.2 Pattern of Non-Degenerate Behaviour

The pattern of ranks for a set of probes is formally defined as a beta distribution with parameters α and β. We use beta distributions to examine the behaviour of non-degenerate probes as introduced in Section 3.1. We assume that there are few degenerate probes in the probe set of a chip. Following from this assumption, if we select probes uniformly at random we would expect that the aggregate pattern of ranks of the selected probes is representative of the majority non-degenerate probes. We expect that the ranks of a set of non-degenerate probes should be uniformly distributed over all ranks from 1 to l.

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 constructed from C as described in Section 2.2. We set $b = 3$ so that hybridizations are partitioned into three blocks corresponding to hybridizations with high $I_j(g)$, hybridizations with low $I_j(g)$ and hybridizations with mid-range $I_j(g)$ for each target $g \in G$. The value of b is not set larger because such a value for b increases the possibility of zero values in the rank count vectors. This is problematic when estimating α and β parameters from the rank count vector. Consider $P' \subseteq P$ a set of randomly chosen probes from chip C. Let $y_{b'}^{P',\mathcal{H}}$ be the rank count vector over the set of all hybridizations $\mathcal{H}$ for $1 \leq b' \leq 3$. Beta distribution parameters are estimated from $y_{b'}^{P',\mathcal{H}}$ and the discretized probability vector $\phi_{\widehat{\alpha}_{b'},\widehat{\beta}_{b'}}$ is calculated from the resulting beta distribution $f_{\widehat{\alpha}_{b'},\widehat{\beta}_{b'}}$ for $1 \leq b' \leq 3$. These discretized probability vectors $\phi_{\widehat{\alpha}_{b'},\widehat{\beta}_{b'}}$ are depicted in Figure 12 and confirm that the ranks of non-degenerate probes are uniformly distributed. This is expected, since the background distribution of ranks (pattern of ranks over all probes) is uniform, as each rank occurs once within each group at every hybridization leading to the same number of occurrences of each rank. We also find that the distributions do not vary greatly over the different blocks. This also indicates that the probes are non-degenerate since the pattern of ranks is consistent over all blocks.

We are now interested in the background distribution of ranks over the set of all probes with similar probe-tag pair free-energy ΔG. We conjecture that the ΔG of a probe-tag pair (p,t), $p \in P$, $t \in T$ measured by $\mathcal{DP}(p,t)$ is correlated with the pattern of ranks displayed by probe p if p is a non-degenerate probe. If $\mathcal{DP}(p,t)$ is low (indicating that (p,t) is a probe-tag pair with strong hybridization strength), probe p should have a tendency to exhibit high ranks. Otherwise, if $\mathcal{DP}(p,t)$ is high, probe p should have a tendency to exhibit low ranks. We test this conjecture by creating

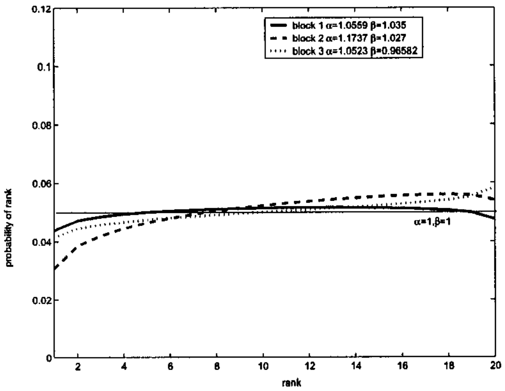


Figure 12: Estimated discretized probability vectors $\phi_{\widehat{\alpha}_{b'},\widehat{\beta}_{b'}}$ where $b = 3$ for a set of probes randomly chosen from P.

randomly selected sets of probes $P' \subseteq P$ such that for probe $p \in P'$ with matching tag t, $\tau_{min} \leq \mathcal{DP}(p,t) \leq \tau_{max}$ for some pair of ΔG thresholds τ_{min} and τ_{max}. Recall from Section 2.2.1 that -42.003 kcal/mol $\leq \mathcal{DP}(p,t) \leq -18.993$ kcal/mol for all probe-tag pairs (p,t), $p \in P$ and $t \in T$. We construct $P_{low} \subseteq P$ to be a randomly selected set of probes such that for all $p \in P_{low}$ with matching tag t, $\mathcal{DP}(p,t) < -34$. $P_{mid} \subseteq P$ and $P_{high} \subseteq P$ are also sets of randomly selected probes such that for $p \in P_{mid}$ with matching tag t, $-34 \leq \mathcal{DP}(p,t) < -26$ and for $p' \in P_{high}$ with matching tag t', $\mathcal{DP}(p',t') \geq -26$. Beta distribution parameters are estimated from rank count vectors $y_{b'}^{P_{low},\mathcal{H}}$, $y_{b'}^{P_{mid},\mathcal{H}}$ and $y_{b'}^{P_{high},\mathcal{H}}$ and discretized probability vectors are calculated from the resulting beta distributions for each of P_{low}, P_{mid} and P_{high}, for $1 \leq b' \leq 3$. Figure 13 depicts the discretized probability vectors of P_{low}, Figure 14 depicts those of P_{mid} and Figure 15 depicts those of P_{high}. We find that the distributions estimated from P_{low} have high density at high ranks, while the distributions estimated from P_{high} have high density at low ranks. The distributions estimated from P_{mid} have similar density over all ranks. These results confirm that ΔG is correlated to the rank pattern of a non-degenerate probe and that the background distributions of ranks of non-degenerate probes within a range of ΔG is different from the background distribution of ranks over all non-degenerate probes. We also find that the beta distributions estimated for the random set of probes in P and for sets of probes P_{low}, P_{mid} and P_{high} (Figures 12 to 15) are consistent with rank patterns of non-degenerate probes since they do not change for $b' = 1$, $b' = 2$ and $b' = 3$.

In the following sections, we determine if the rank count vector of a probe p is more likely to be sampled from the background distribution of ranks of a non-degenerate probe or to be sampled from a distribution of ranks corresponding to a degenerate behaviour. For this purpose, we require the most accurate estimation of the background distribution of a non-degenerate probe. To this end we use the background distribution of a non-degenerate probe within a particular range of ΔG values.

Definition 9 (Background Beta Distribution) *Let* $y_{b'}^{P',\mathcal{H}}$ *be the rank count vector of* $P' = \{p \in P$ *with matching tag* $t \in T : \tau_{min} \leq \mathcal{DP}(p,t) \leq \tau_{max}\}$ *for some pair of* ΔG *thresholds* τ_{min} *and* τ_{max}, $1 \leq b' \leq b$. $f_{\widehat{\alpha}_{b'},\widehat{\beta}_{b'}}$ *is the beta distribution with parameters* $\widehat{\alpha}_{b'}$ *and* $\widehat{\beta}_{b'}$ *estimated from* $y_{b'}^{P',\mathcal{H}}$. *We define* $f_{b'}(\tau_{min},\tau_{max}) = f_{\widehat{\alpha}_{b'},\widehat{\beta}_{b'}}$ *to be the* background beta distribution *of a non-degenerate probe* $p \in P$ *with matching tag* $t \in T$ *with* $\tau_{min} \leq \mathcal{DP}(p,t) \leq \tau_{max}$ *at block* b'.

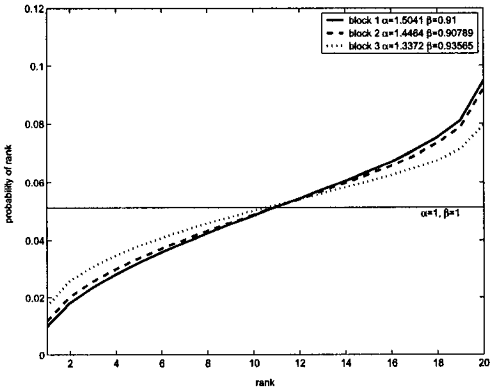


Figure 13: Discretized probability vectors $\phi_{\hat{\alpha},\hat{\beta}}$ calculated from rank count vectors $y_{b'}^{P_{low},\mathcal{H}}$ for set of randomly selected probes $P_{low} \subseteq P$ such that for $p \in P_{low}$ with matching tag t, $\mathcal{DP}(p,t) < -34$.

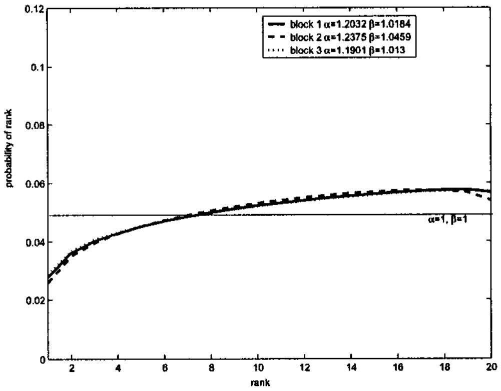


Figure 14: Discretized probability vectors $\phi_{\hat{\alpha},\hat{\beta}}$ calculated from rank count vectors $y_{b'}^{P_{mid},\mathcal{H}}$ for set of randomly selected probes $P_{mid} \subseteq P$ such that for $p \in P_{mid}$ with matching tag t, $-34 \leq \mathcal{DP}(p,t) < -26$.

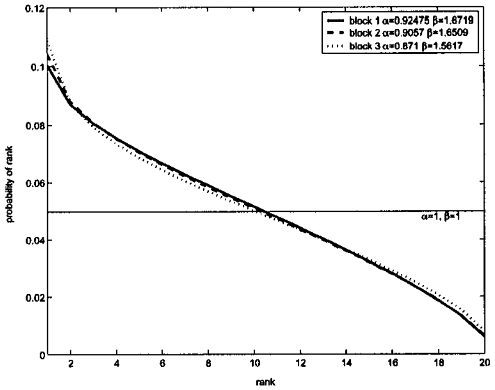


Figure 15: Discretized probability vectors $\phi_{\hat{\alpha},\hat{\beta}}$ calculated from rank count vectors $y_{b'}^{P_{high},\mathcal{H}}$ for set of randomly selected probes $P_{high} \subseteq P$ such that for $p \in P_{high}$ with matching tag t, $\mathcal{DP}(p,t) \geq -26$.

3.3 Pattern of Degenerate Behaviour

We now examine the pattern of ranks of degenerate probes. We provide an intuitive hypothesis of the pattern of ranks of a probe affected by each one of the four degeneracies. We justify the conjectured patterns by showing their similarity to the patterns of ranks of probes represented by endpoints of edges in the edge sets S, SD, X, D of conflict graph M from Section 2.1 (probes measured to have a high affinity for degeneracy).

3.3.1 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. Firstly, 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 non-degenerate tags hybridize with members of P_g. Therefore, the rank of p in this experiment, $\rho_j(p)$, is expected to be very low. Secondly, if the target g is lowly expresed 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. We bin the intensity values for all genes 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 $1 \leq i \leq b$ depicted in Figure 16 for $b = 3$ re-affirm the description of secondary structure forming probes. The probability vectors of S' confirm the intuition that probes predicted to form secondary structure behave similar to non-degenerate probes when the target is lowly expressed ($b = 1$) since probes in S' have low probe-tag pair ΔG and they would be expected to have background beta distributions similar to that of P_{low} depicted in Figure 13. In fact, we find that probes in S' exhibit fewer high ranks than would be expected for non-degenerate probes within the same probe-tag pair ΔG range. As the target becomes highly expressed ($b = 3$), the intuition for secondary structure behaviour is again confirmed since the probes in S' have high probability of very low ranks and low probability of very high ranks. More formally, $\phi_{\widehat{\alpha}_1,\widehat{\beta}_1}$ has a near 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.

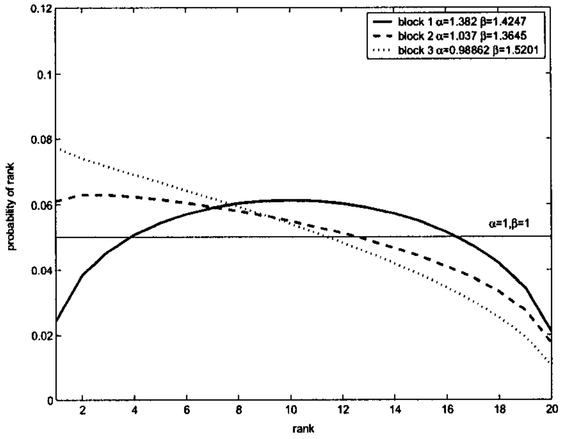


Figure 16: Estimated discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \leq i \leq b$ for the set of probes S' predicted to have an affinity to form secondary structure for $\tau_s = -6$, $b = 3$.

3.3.2 Self-dimerization

We conjecture that probes with an affinity to self-dimerize will follow the same behavioural conjecture as that used for secondary structure. We tested 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, $1 \leq i \leq b$. As for secondary structure, we use Equation 1 to compute estimate parameters $\widehat{\alpha}_i$, $\widehat{\beta}_i$ for a beta distribution. Figure 17 depicts the discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ calculated from the resulting $f_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \leq i \leq 3$. Our intuition of the pattern of ranks of self-dimerizing probes is re-affirmed by the discretized probability vectors. We find that ranks of probes in SD' in block 1 are distributed as non-degenerate probes with low probe-tag pair ΔG ($-35.213 \leq \mathcal{DP}(p,t) \leq -28.613$ for $p \in SD'$ with matching tag t). And as the target becomes highly expressed ($b = 3$) there is a decrease in the probability of high ranks and an increase in the probability of low ranks. Note that this re-affirmation for self-dimerization is not as strong as that for secondary structure.

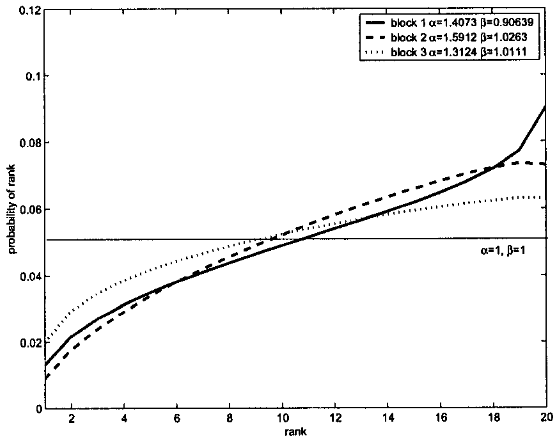


Figure 17: Estimated discretized probability vectors $\phi_{\hat{\alpha}_i,\hat{\beta}_i}$ for $1 \leq i \leq b$ for the probes belonging to the set of predicted probes prone to self-dimerization SD' for $\tau_{sd} = -14$, $b = 3$.

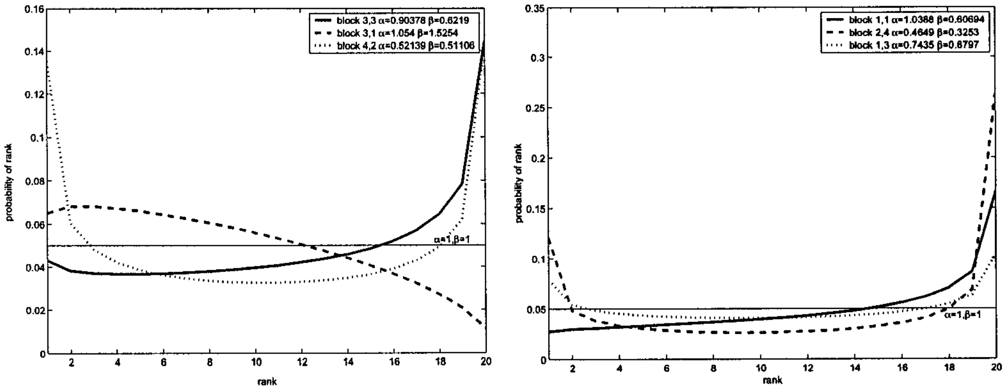
3.3.3 Cross-Hybridization

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 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 $P_{g'}$. Simlarily, 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 non-degenerate 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 genes 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, $1 \leq i, j \leq b$. Figure 18 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 $1 \leq i, j \leq b$ where $b = 4$. The discretized probability vector $\phi_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$ is labelled by "block i,j". Each probability $\phi_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}[k]$, for $1 \leq k \leq l$, is the probability that probe $p \in X'$ is a rank k occurrence when p is in block i and probe p' is in block j. The discretized probability vectors confirm the intuition regarding cross-hybridization described above.

Figure 18(a) depicts block pairs (values of i, j) when i is fixed at a high value ($i = 3$ or $i = 4$) and the value of j varies. Block pairs $(3, 1)$ and $(4, 2)$ are hybridizations when p is expected to lose tags to p' since g is highly expressed and g' is lowly expressed. Similarly, Figure 18(b) depicts block pairs when i is fixed at a low value ($i = 1$ or $i = 2$) and the value of j varies. Block pairs $(2, 4)$ and $(1, 3)$ are hybridizations when p is expected to gain tags from p' since g is lowly expressed and g' is highly expressed.

When $i = 3$ and $j = 1$ and slightly less when $i = 4$ and $j = 2$, we observe that the probability of high ranks decreases and there is an increase in the probability of low or middle ranks. When $i = 2$ and $j = 4$, we observe an increase in the probability of high ranks, though we also observe an increase in the probability of low ranks. It may be the case that some probes p do indeed gain tags from p', while others do not and continue to have a high probability of low ranks. When $i = 1$ and $j = 3$, we do not observe a curve indicating cross-hybridization. We note that these distributions for cross-hybridization do not re-affirm our intuition as strongly as in the cases of secondary structure and self-dimerization. The estimated parameters $\widehat{\alpha}_{i,j}, \widehat{\beta}_{i,j}$, $1 \leq i, j \leq b$ for all b^2 block pairs are available online [21]. Block pairs $(4, 1)$ and $(1, 4)$ are not included because there were no such pairwise occurrences in the data sets.



(a) (b)

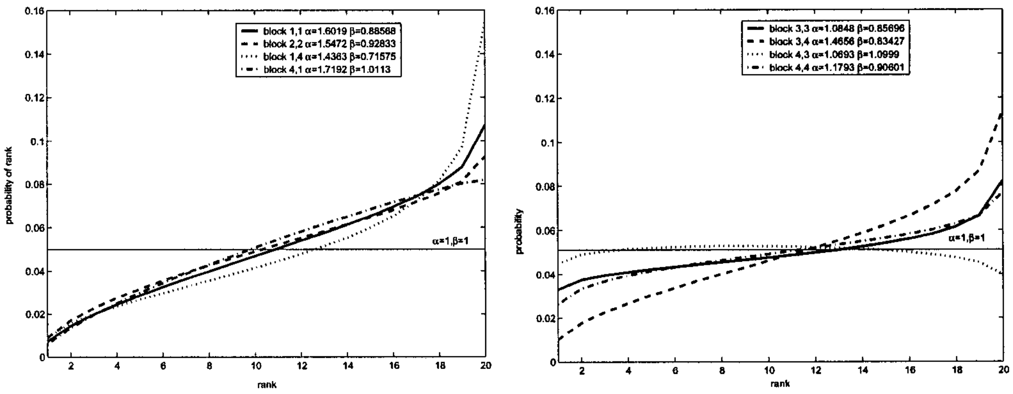
Figure 18: Estimated discretized probability vectors $\phi_{\widehat{\alpha},\widehat{\beta}}$ for the set of probe pairs X' predicted to have an affinity to cross-hybridize with $\tau_x = -23$ and $b = 4$. Figure (a) depicts block pairs (values of i, j) when i is fixed at a high value ($i \approx b$) and Figure (b) depicts block pairs when i is fixed at a low value ($i \approx 1$).

3.3.4 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 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 non-degenerate probes.

Let $D' = \{p, p' : (p, p') \in D\}$ be probes predicted to exhibit dimerization. We bin the intensity values for each of the genes g into b blocks (Definition 3). Let $y_{(i,j)} = y^{D',H}_{(i,j)}$ 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, $1 \leq i, j \leq b$. Figure 19 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 $1 \leq i, j \leq b$ where $b = 4$. As with cross-hybridization, discretized probability vector $\phi_{\hat{\alpha}_{i,j}, \hat{\beta}_{i,j}}$ is labelled "block i,j" and each probability in $\phi_{\hat{\alpha}_{i,j}, \hat{\beta}_{i,j}}[k]$ for $1 \leq k \leq l$ is the estimated probability that probe $p \in D'$ is a rank k occurrence when p is in block i and its dimerizing pair p' is in block j.

Figure 19(a) depicts block pairs when dimerization should not affect the behaviour of the probes (when $i \approx j \approx 1$) while Figure 19(b) depicts block pairs when dimerization should raise the probability of low ranks (when $i \approx j \approx 1$). We found that in block pairs where $i \approx j \approx b$, the estimated probability vector $\phi_{\widehat{\alpha}_{(i,j)}, \widehat{\beta}_{(i,j)}}$ indicates that it is likely that probes in D' have lower ranks. And when $i \approx j \approx 1$ or $i \approx 1$ and $j \approx b$ (and $j \approx 1$ and $i \approx b$), the estimated probability vectors indicate a non-degenerate behaviour of probe p. This is true since as with X', probes in D' are measured to have low probe-tag pair ΔG and thus if they behaved in a non-degenerate manner, their background distribution would be similar to that of probes set P_low from Section 3.2. Thus, the discretized probability vectors $\phi_{\hat{\alpha}_{i,j},\hat{\beta}_{i,j}}$ re-affirm the intuition regarding dimerization described in Section 3. 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 online [21].



(a) (b)

Figure 19: 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 $b = 4$. In Figure (a) block pairs where dimerization does not occur are depicted an in Figure (b) block pairs where dimerization occurs are depicted.

Note that for pairwise degeneracies we use $b = 4$, since we require a large difference between the intensities of the two targets to reflect situations when one target is highly expressed and the other is lowly expressed.

4 Experimental Support

In Section 3, we show that it is possible to define unique patterns for each type of degeneracy and that beta distributions can formally define each such pattern. We also show that the background distribution of ranks of a set of non-degenerate probes can be defined as a beta distribution. The discretized probability vector $\phi_{\alpha,\beta}$ derived from beta distribution $f_{\alpha,\beta}$ associated with a degenerate behaviour gives us a probability vector $p = \langle p_1, \ldots, p_l \rangle$ where p_i is the probability that a degenerate probe p' is a rank i occurrence at any hybridization, $1 \leq i \leq l$. We now witness a rank count vector $y = \langle y_1, \ldots, y_l \rangle$ from a degenerate probe p' and we want to know the probability of y, given the probability vector p. In such a case, the probability of the rank count vector y of probe p' is given by the *multinomial distribution formula* defined as follows. Let $X_1, \ldots, X_n$ be a sequence of independent identically distributed random variables each taking one of l possible values $1, \ldots, l$. Values occur with fixed probabilities $p = \langle p_1, \ldots, p_l \rangle$. Let $Y_1, \ldots, Y_l$ be random variables that count the number of times each of the l values occur in $X_1, \ldots, X_n$. The probability that $Y[i] = y[i]$ for $1 \leq i \leq l$ is given by multinomial distribution formula

$$P_{Y,p}(y) = \frac{n!}{\prod_i (y[i]!)} \prod_i p[i]^{y[i]}.$$

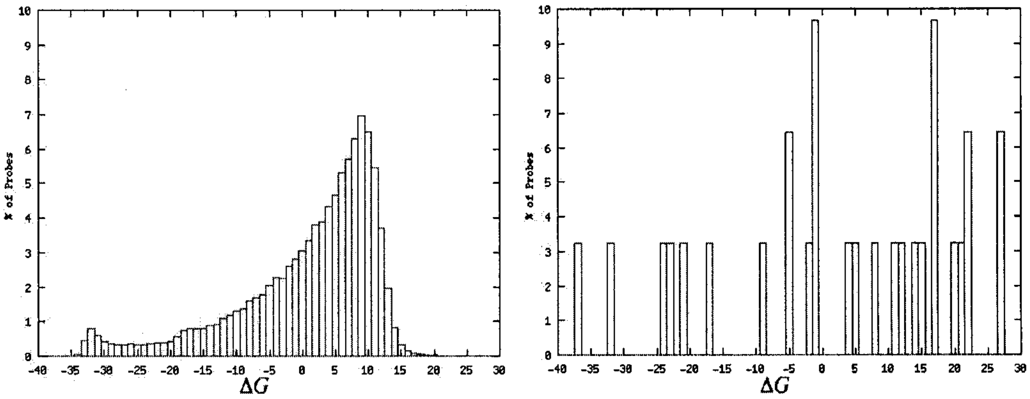
The multinomial distribution formula is used in our *experimental support functions*, described below for each degenerate behaviour. Each experimental support function measures the support from a set of hybridizations that the ranks of a probe $p' \in P$ are distributed according to a discrete probability distribution that has been associated with a degenerate behaviour. We use the experimental support functions to determine if a rank count vector of a probe p supports or contests the claim that p is degenerate.

Firstly, consider the secondary structure formation degenerate behaviour. The experimental secondary structure support function $\widehat{\mathcal{S}}(p)$ is the average, over all blocks, of the log ratio of the probability of observing the rank count vector given that probe p is prone to secondary structure and of the probability of seeing the rank count vector given that p is non-degenerate. Let ϕ_i, $1 \leq i \leq b$, be the discretized probability vector derived from a beta distribution associated with secondary structure behaviour. In our experiments, ϕ_i is derived from the estimated beta distribution for secondary structure $f_{\widehat{\alpha}_i,\widehat{\beta}_i}$ described in Section 3.3.1. Let ϕ'_i, $1 \leq i \leq b$ be the discretized probability vector for block i derived from the background beta distribution $f_i(\tau_{min}, \tau_{max})$ of ranks of probe p, given that p is non-degenerate, as defined in Section 3.2. The ΔG thresholds are $\tau_{min} \leq \mathcal{DP}(p,t) \leq \tau_{max}$ for matching tag $t \in T$ such that $f_i(\tau_{min}, \tau_{max})$ is estimated from a sufficiently large set of probes. Given the collection of rank count vectors $y = \langle y_1 \ldots y_b \rangle$ for probe p over hyridization set $\mathcal{H}$, let

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

We examine the ability $\widehat{\mathcal{S}}$ to measure the experimental support that a probe is degenerate due to secondary structure formation. Recall that S is the set of all s labelled edges of conflict graph M as defined in Section 2.1. Let $S' = \{p \in V(G) : (p,p) \in S\}$ be the set of probes predicted to exhibit secondary structure. We calculate $\widehat{\mathcal{S}}(p)$ for all probes $p \in P$.

Figure 20(a) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{S}}(p)$ for all $p \in P$ and Figure 20(b) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{S}}(p)$ for all $p \in S'$ for $b = 3$. We find that $\widehat{\mathcal{S}}$ does not indicate that the experimental data supports the claim that all probes $p' \in S'$ are degenerate due to secondary structure formation. We would expect that $\widehat{\mathcal{S}}(p)$ for all $p \in S'$ are among the highest over the range of values calculated by $\widehat{\mathcal{S}}$. However, we find that that $\widehat{\mathcal{S}}$ does not discriminate between probes in S' and probes in $P \backslash S'$.



(a) (b)

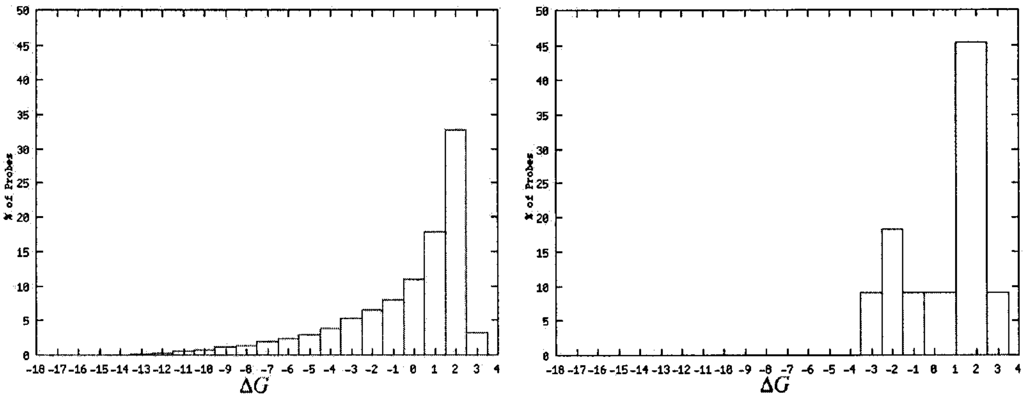
Figure 20: Depicted here are histograms of secondary structure formation support values measured by $\widehat{\mathcal{S}}$ for (a) the set of probes P and (b) the set of probes S'.

The experimental self-dimerization support function $\widehat{\mathcal{SD}}(p)$ defined as the average, over all blocks, of the log ratio of the probability of observing the rank count vector given that probe p is prone to self-dimerization and of the probability of seeing the rank count vector given that p is non-degenerate. Function $\widehat{\mathcal{SD}}(p)$ measures the level to which hybridization data supports the claim that probe p is degenerate due to self-dimerization. Let ϕ_i, $1 \le i \le b$, be the discretized probability vector of a beta distribution associated with self-dimerization. In our experiments, ϕ_i is derived from the estimated beta distribution of self-dimerization $f_{\widehat{\alpha}_i, \widehat{\beta}_i}$. Let ϕ'_i, $1 \le i \le b$ be the discretized probability vector for block i derived from the background beta distribution $f_i(\tau_{min}, \tau_{max})$ of ranks of probe p, given that p is non-degenerate, as defined in Section 3.2. The ΔG thresholds are $\tau_{min} \leq \mathcal{DP}(p,t) \leq \tau_{max}$ for matching tag $t \in T$ such that $f_i(\tau_{min}, \tau_{max})$ is estimated from a sufficiently large set of probes. Given the collection of rank count vectors of p over hyridization set $\mathcal{H}$, $y = \langle y_1, \dots y_b, \rangle$ let

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

As with $\widehat{\mathcal{S}}$, we wish to examine the ability of $\widehat{\mathcal{SD}}$ to measure the experimental support that a probe is degenerate due to self-dimerization. We calculate $\widehat{\mathcal{SD}}(p)$ for all probes $p \in P$. Let $SD' = \{p \in V(G) : (p,p) \in SD\}$ be the set of probes predicted to exhibit self-dimerization in conflict graph M. Figure 21(a) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{SD}}(p)$ for all $p \in P$ and Figure 20(b) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{SD}}(p)$ for all $p \in SD'$ for $b = 3$. We find that $\widehat{\mathcal{SD}}$ does not discriminate between probes in SD' and probes in $P \backslash SD'$ since values $\widehat{\mathcal{SD}}(p)$ for $p \in SD'$ are consistent with a set of values randomly sampled from values of $\widehat{\mathcal{SD}}$ of all probes in P.

We measure the support that a pair of probes (p, p') display cross-hybridization behaviour in hybridization set $\mathcal{H}$ with support function $\widehat{\mathcal{X}}(p,p')$. Let $\phi_{i,j}$, $1 \leq i,j \leq b$, be the discretized probability vector derived from a beta distribution associated with cross-hybridization. In our experiments, $\phi_{i,j}$ is derived from $f_{\widehat{\alpha}_{(i,j)}, \widehat{\beta}_{(i,j)}}$, $1 \leq i,j \leq b$. Let any undefined $\phi_{i,j}$ be equal to vector consistent with the pattern of cross-hybridization. Let $\phi'_{i,j}$, $1 \leq i,j \leq b$, be the discretized probability vector for block i derived from the background beta distribution $f_i(\tau_{min}, \tau_{max})$ of ranks of probes p



(a)

(b)

Figure 21: Depicted here are histograms of self-dimerization support values measured by $\widehat{\mathcal{SD}}$ for (a) the set of probes P and (b) the set of probes SD'.

and p', given that p and p' are non-degenerate probes. The ΔG thresholds are $\tau_{min} \leq min\ \{\mathcal{DP}(p,t), \mathcal{DP}(p',t')\}$ and $\tau_{max} \geq max\ \{\mathcal{DP}(p,t), \mathcal{DP}(p',t')\}$ for matching tags $t, t' \in T$ such that $f_i(\tau_{min}, \tau_{max})$ is estimated from a sufficiently large set of probes. Given the collection of pairwise rank count vectors $y = \langle y_{1,1} \dots y_{b,b} \rangle$ for probe pair (p,p') over hybridization set $\mathcal{H}$, let

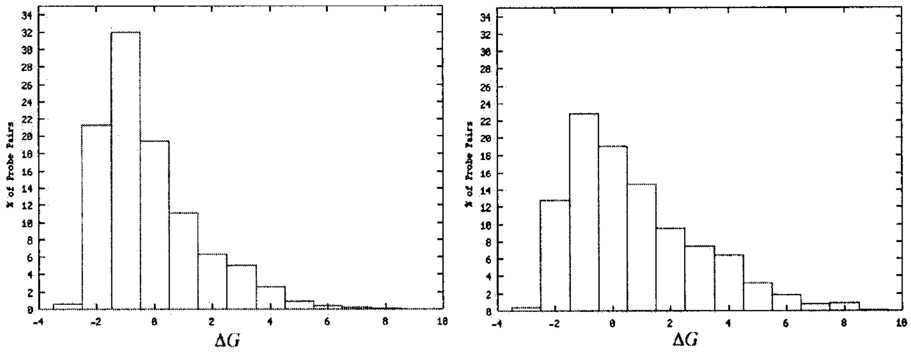
$$\widehat{\mathcal{X}}(p,p') = \frac{1}{b^2} \sum_{1 \leq i,j \leq b} log \left(\frac{P_{Y_{(i,j)},\phi_{(i,j)}}(y_{(i,j)})}{P_{Y_{(i,j)},\phi'_{(i,j)}}(y_{(i,j)})} \right).$$

Recall that set X contains pairs of probes (p,p') predicted by conflict graph M to exhibit cross-hybridization. We calculate $\widehat{\mathcal{X}}(p,p')$ for all probe pairs $(p,p') \in X$ and for 1.7 million randomly selected probe pairs sampled from the eight billion pairs in the population[2]. Figure 22(a) depicts a histogram of the percentage of support values

---

[2] The results on the random selection of probe pairs reflect the total probe pair population with a confidence level of 99% and a confidence interval of 0.1.

measured by $\widehat{\mathcal{X}}(p,p')$ over the set of randomly chosen probe pairs and Figure 22(b) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{X}}(p,p')$ for all $(p,p') \in X$ for $b = 4$. We find that $\widehat{\mathcal{X}}$ is unable to discriminate between probe pairs $(p,p') \in X$ and probe pairs $(p,p') \notin X$. This is clear since we would expect that support values of $\widehat{\mathcal{X}}(p,p')$ for $(p,p') \in X$ are among the highest support values $\widehat{\mathcal{X}}$ over all probe pairs. Instead, we find that the support values $\widehat{\mathcal{X}}(p,p')$ for $(p,p') \in X$ follow the underlying distribution $\widehat{\mathcal{X}}$ over all probe pairs.



(a)

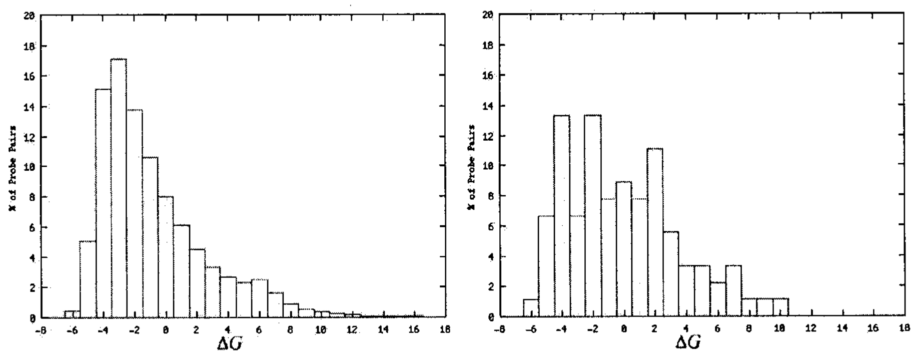
(b)

Figure 22: Figure (a) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{X}}(p,p')$ for a large randomly selected set of probe pairs $(p,p') \in P \times P$ and (b) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{X}}(p,p')$ for $(p,p') \in X$.

The experimental support function for dimerization $\widehat{\mathcal{D}}$ is very similar to $\widehat{\mathcal{X}}$ in both definition and its ability to discriminate dimerizing probe pairs. Support function $\widehat{\mathcal{D}}$ for a probe pair (p,p') is defined as follows:

$$\widehat{\mathcal{D}}(p,p') = \frac{1}{b^2} \sum_{1 \leq i,j, \leq b} log \left(\frac{P_{Y_{(i,j)},\phi_{(i,j)}}(y_{(i,j)})}{P_{Y_{(i,j)},\phi'_{(i,j)}}(y_{(i,j)})} \right).$$

Where $\phi_{i,j}$, $1 \leq i,j \leq b$ is defined as for $\widehat{\mathcal{X}}$ but is derived from a beta distribution associated with dimerization. D contains pairs of probes (p,p') predicted by conflict graph M to exhibit dimerization. We find that $\widehat{\mathcal{D}}$ is unable to discriminate between probe pairs $(p,p') \in D$ and probe pairs $(p,p') \notin D$. Figure 23 depicts histograms of the percentage of support values measured by $\widehat{\mathcal{D}}$ for (a) a set of 1.7 million randomly selected probe pairs from $P \times P$ and (b) the set of probe pairs $(p,p') \in D$.



(a)

(b)

Figure 23: Figure (a) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{D}}(p,p')$ for a large randomly selected set of probe pairs $(p,p') \in P \times P$ and (b) depicts a histogram of the percentage of support values measured by $\widehat{\mathcal{D}}(p,p')$ for $(p,p') \in D$.

We found that in HuGeneFL, the experimental support functions were not able to detect a pattern of degeneracy in the hybridization data. It is expected that the experimental support functions should assign large support values to the members of sets of probes (or probe pairs) predicted to be degenerate by conflict graph M. However, the support values measured for probes (and probes pairs) in these probes sets behave at random and are not reflective of probe sets that are known to exhibit patterns of degeneracy in the hybridization data.

Another indication of the failure of the experimental support functions is the distribution of experimental support values for all four types of degeneracy. The majority of probes are measured by $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{SD}}$ have positive support values, meaning that hybridization data supports the claim that the majority of probes are degenerate. This directly contradicts the assumption that there are few degenerate probes in the probe set P. We find the same contradiction in the support values measured by $\widehat{\mathcal{X}}$ and $\widehat{\mathcal{D}}$ of a randomly selected set of 1.7 million probe pairs sampled from the eight billion pairs in the population.

4.1 Reasons for Failure of Experimental Support Functions

We believe that counting the frequency of ranks is a very good method for determining degenerate behaviour of probes over a set of hybridization experiments. However, the experimental support functions used to detect the degeneracy from the observed frequencies did not succeed. We believe that the experimental support functions did not succeed when considering a solitary probe (or probe pair) at a give block (or block pair) because there too few hybridizations within the block (or block pair) to ensure that each of the rank categories has a sufficiently large frequency. This occurs because the experiment partitions an already small hybridization set into b blocks (often resulting in as few as 1 and 5 hybridizations within a block). In addition, there are many rank categories (20), leading to high instances of sampling zeroes, cases when the count of a rank is zero.

The experimental support functions use a log likelihood goodness-of-fit test. We designed experimental support functions with several other different goodness-of-fit tests without success. Consider a chi-square test. Such a test works best with a frequency of at least 5 in each category (in this model the category is rank) [11]. We apply this rule to our experimental model to determine an estimate on the minimum required size of hybridization set $\mathcal{H}$. In our estimated probability distributions of degenerate probes and non-degenerate probes, the probability of a rank may be as low as 0.01. Therefore 500 hybridizations per block would be a very liberal estimate on the required number of hybridizations. Recall that the entire set of hybridizations must be partitioned into b blocks (or b^2 block pairs). Although b can be chosen according to the total number of hybridizations, it should not be smaller than three. If the hybridizations were uniformly distributed over the blocks, then we would required 1500 hybridizations ($|\mathcal{H}| = 1500$) in the sample to guarantee that the statistical test is applied to enough data. A much larger hybridization set is required for the pairwise tests.

We conjecture that the main reason for failure of the experimental support functions is the small size of $\mathcal{H}$ used in our experiments. We formalize the following questions.

> *Question*: What is the total minimum size of rank count vector and the minimum value for each category in the rank count vector required to

guarantee that the α, β parameter estimates define a beta distribution with very good fit to the rank count vector?

Question: What is the total minimum size of rank count vector and the minimum value for each category in the rank count vector required for a successful goodness-of-fit test.

The answers to these questions will allow us to determine the minimum required number of hybridizations required and will help us to choose the proper value of b, the number of blocks of hybridizations.

5 Learning from Hybridization Data

One strength of this framework lies in its potential ability to learn suspect probe-tag pairs from hybridization data. If we can accurately predict the behavior of a probe with no prior knowledge regarding the conditions of hybridizations with the chip, then this probe adds no new information to the experiment and should be removed from the chip. Suppose that the experimental support functions defined in Section 4 did have the ability to provide a method to discriminate between the amount that experimental hybridization data supports a claim of degeneracy. With these experimental support functions we could extend the conflict graph to enable the identification of degenerate patterns in the large amount of hybridization data that would indicate the presence of a degenerate probe. Although the experimental support functions do not behave as desired, we include the following description of a learning technique from hybridization data so that it may be used once successful experimental support functions are determined.

5.1 Extended Conflict Graph of a Chip

Let $C = (T, A)$ be a chip of interest and let $M = (V, E, \tau, \kappa)$ be constructed as described in Section 2. The κ function labelling the edges of M is extended to include experimental edge labels $\kappa : E \rightarrow \{s, x, d, \hat{s}, \hat{x}, \hat{d}\}$, and $\{\tau_{\widehat{s}}, \tau_{\widehat{sd}}, \tau_{\widehat{x}}, \tau_{\widehat{d}}\}$ are suitable threshold parameters for the $\widehat{\mathcal{S}}$, $\widehat{\mathcal{SD}}$, $\widehat{\mathcal{X}}$ and $\widehat{\mathcal{D}}$ experimental support functions defined in Section 4. The K hybridization experiments $\mathcal{H} = \{H_1, \ldots, H_K\}$ are randomly partitioned into two halves called the *training set* and *verification set.* W.l.o.g. let the training set hybridizations be $H_T = \{H_1, \ldots, H_{\lfloor K/2 \rfloor = k}\}$.

The addition of *experimental edges* to M is accomplished by examining the probe intensities in the hybridization experiments of the training set with experimental support functions $\widehat{\mathcal{S}}$, $\widehat{\mathcal{SD}}$, $\widehat{\mathcal{X}}$ and $\widehat{\mathcal{D}}$. Using $\widehat{\mathcal{S}}$, a self-loop (p, p) with $\kappa(p, p) \leftarrow \hat{s}$ is added to the conflict graph M iff $\widehat{\mathcal{S}}(p, p) > \tau_{\hat{s}}$. Similarly, a self-loop (p, p) with $\kappa(p, p) \leftarrow \hat{d}$ is added to the conflict graph G iff $\widehat{\mathcal{SD}}(p, p) > \tau_{\widehat{sd}}$. An edge (p, p'), $p \neq p'$ is included in G iff $\widehat{\mathcal{X}}(p, p') > \tau_{\hat{x}}$ (or $\widehat{\mathcal{D}}(p, p') > \tau_{\hat{d}}$). The set of all *experimental secondary structure, cross-hybridization*, or *dimerization* edges respectively are denoted $\widehat{S}, \widehat{X}$ and $\widehat{D}$.

Through building M we predict the behaviour of each probe and probe pair in chip C. Probes predicted to exhibit a degenerate behaviour are organized into sets of the same type. Let $S = \{(p, p) \in E \text{ and } \kappa(p, p) = s\}$ and $\widehat{S} = \{(p, p) \in E \text{ and } \kappa(p, p) = \hat{s}\}$. Let $SD = \{(p, p) \in E \text{ and } \kappa(p, p) = d\}$ and $\widehat{SD} = \{(p, p) \in E \text{ and } \kappa(p, p) = \hat{d}\}$. Let $X = \{(p, p') \in E \text{ and } \kappa(p, p') = x \text{ and } p \neq p'\}$. Define $\widehat{X}$, D, and $\widehat{D}$ similarily. Probes in S, SD, X or D are predicted by theoretical supports to exhibit the respective degeneracy and probes in $\widehat{S}, \widehat{SD}, \widehat{X}$ or $\widehat{D}$ are predicted by experimental supports to exhibit the respective degeneracy.

A truly degenerate probe exhibits a particular pattern over all hybridizations. We test the ability of M to detect degenerate probes or probe pairs by testing if a predicted degenerate probe exhibits the expected pattern in subsequent (unseen) hybridizations. Our verification set H_V consists of the remaining hybridization experiments $\mathcal{H} \backslash H_T$. The accuracy of the ability of this method to detect probes affected by secondary structure is measured by considering each probe p for which $(p,p) \in S$ or $(p,p) \in \widehat{S}$. We ask whether the ranks of p over H_V support the decision that p has the affinity to form secondary structure. The answer is given by comparing the probability that p follows the rank distribution of a secondary structure probe to the probability that p follows the rank distribution of a well behaved probe. This ratio is calculated with support function $\widehat{S}$.

$$Score_S(S) = \frac{1}{|S|} \cdot \sum_{(p,p) \in S} \begin{cases} 1 & \text{if } \widehat{S}(p) > \tau_{\hat{s}} \text{ for } y^{p,H_V} = y_1^{p,H_V} \dots y_b^{p,H_V}. \\ 0 & \text{otherwise.} \end{cases}$$

The prediction accuracy of $\widehat{S}$, $Score_S(\widehat{S})$ is calculated similarly. It is possible to test if the secondary structure detection ability of M is improved by combining the theoretical and experimental detection approaches by scoring $S \cap \widehat{S}$. We also must consider the specificity of the predicted set. To do so, we count the number of probes p such that $\widehat{S}(p) > \tau_{\hat{s}}$ for $y^{p,H_V} = y_1^{p,H_V} \dots y_b^{p,H_V}$ and $(p,p) \notin S$ or $(p,p) \notin \widehat{S}$. $Score_{SD}$ is defined similarly using $\widehat{SD}$ and the same series of questions are posed for prediction sets SD, $\widehat{SD}$ and $SD \cap \widehat{SD}$.

The same approach is used to measure the accuracy and specificity of the cross-hybridization prediction sets. Consider the accuracy of set X,

$$Score_X(X) = \frac{1}{|X|} \cdot \sum_{(p,p') \in X} \begin{cases} 1 & \text{if } \widehat{\mathcal{X}}(p,p') > \tau_{\hat{x}} \text{ for } y^{(p,p'),H_V} = y_{(1,1)}^{(p,p'),H_V} \dots y_{(b,b)}^{(p,p'),H_V}. \\ 0 & \text{otherwise.} \end{cases}$$

for support function $\widehat{\mathcal{X}}$. Accuracy scores $Score_X(\widehat{(X)})$ and $Score_X(X \cap \widehat{(X)})$ are also calculated. $Score_D$ is defined for $\widehat{\mathcal{D}}$ similarly.

For all score functions, higher scores are better indicators of detection power than near zero scores.

Due to the inability of the experimental support values to properly measure the degeneracy or non-degeneracy of a probe or probe pair based solely on hybridization data, it is impossible to use appropriate experimental thresholds to reject or accept the hypothesis of degeneracy using the experimental support functions. Therefore, we can not add experimental edges to M to construct sets $\widehat{S}, \widehat{SD}, \widehat{X}$ and $\widehat{D}$ or calculate prediction scores.

We found that problems with the experimental support functions and their application to the HuGeneFL data set are apparent in the learning method since there is an absence of variation between support values calculated over H_T and H_V. The support values calculated over H_T for a probe or probe pair are nearly identical to those calculated of the value over H_V. The same is true for all three remaining types of degeneracy. We would expect that the experimental support functions applied to different hybridization sets should lead to more variation in the support values.

6 Affymetrix Discrimination

As we stated in the introduction, the framework and methods described in this paper can also be used to investigate additional properties of the probe set of a chip. We are interested in examining the *discrimination* property of each probe on an Affymetrix GeneChip. The discrimination property measures the ability of the intensity of a probe to reflect the true amount of target mRNA transcripts in the sample solution. Affymetrix describes a chip as a set of probe pairs comprised of *Perfect Match (PM)* and *Mismatch (MM)* probe cells. Each PM probe is a probe as we have defined it throughout this paper, though now there is an MM probe associated with each probe. The MM probe of a probe p is a probe with a sequence that differs from the sequence of p by one base pair only. Recall that at each hybridization j in the hybridization set $\mathcal{H}$, $I_j(p)$ is the intensity of probe $p \in P$ of chip C. We let $\overline{I_j(p)}$ be the intensity of the mismatch probe of p at hybridization j.

The expression analysis of Affymetrix GeneChips is performed by a statistical detection algorithm that measures the presence or absence of a transcript. The detection algorithm is a voting algorithm that combines votes from each probe of a probe group to assign a call of present, marginal or absent to the transcript targeted by the probe group. The vote of each probe at hybridization $j \in \mathcal{H}$ is its discrimination score $R_j(p)$ defined as follows:

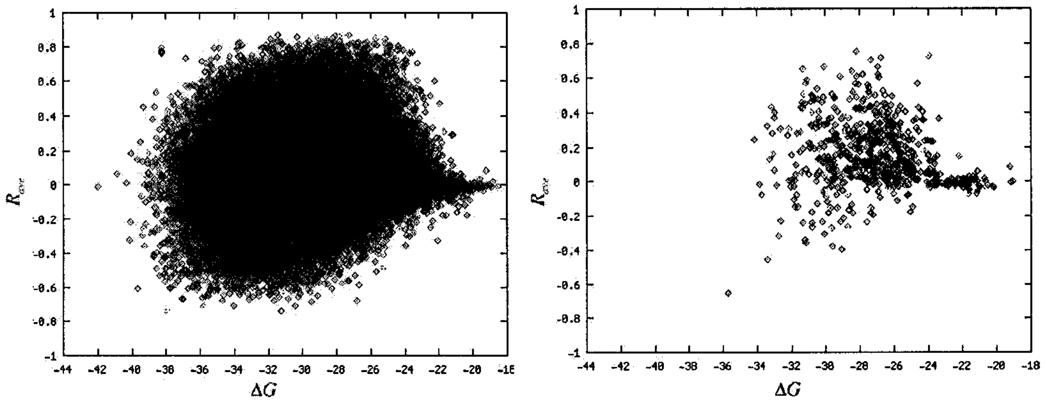
$$R_j(p) = \frac{I_j(p) - \overline{I_j(p)}}{I_j(p) + \overline{I_j(p)}}$$

The discrimination of a probe $R_j(p)$ meaures the ability of the probe to detect its intended target by taking the target-specific intensity difference $(I_j(p) - \overline{I_j(p)})$ relative to the overall hybridization intensity $(I_j(p) + \overline{I_j(p)})$. 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 more 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 less 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.

Consider conflict graph M for HuGeneFL. Let $R_j(p)$ be the discrimination of each probe $p \in P$ at hybridization $j \in \mathcal{H}$ and let $\overline{R(p)} = \frac{1}{|\mathcal{H}|} \cdot \sum_{j \in \mathcal{H}} R_j(p)$ 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) (a measure of the hybridization strength of (p, t). Figure 24(a) depicts a scatterplot of this $\overline{R(p)}$ and ΔG comparison. We find that probes $p \in P$ with high ΔG ($\mathcal{DP}(p, t) > -22$) for matching tag $t \in T$, unanimously have $\overline{R(p)} \approx 0$. This indicates that several probe-tag pairs of HuGeneFL are too weak to discriminate between expressed and non-expressed states.

It is possible that a near zero average discrimination of a probe is caused by the expression level of the target of the probe. To test this, we focus only on probes in probe groups P_g such that there exists a probe $p' \in P_g$ with $\mathcal{DP}(p', t') > -22$ for matching tag $t' \in T$. Figure 24(b) depicts $\overline{R(p)}$ of all probes $p \in P$ such that $p \in P_g$ and there exists a probe $p' \in P_g$ with $\mathcal{DP}(p', t') > -22$ for matching tag $t' \in T$. We find that the near zero average discrimination of probe $p' \in P_g$ is not solely caused by the expression level of target $g \in G$ since $\overline{R(p)}$ takes values both much above and below zero for $p \in P_g$ where $\mathcal{DP}(p, t) \leq -22$ for matching tag t.

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



(a) (b)

Figure 24: Figure (a) depicts a scatterplot of $\overline{R(p)}$ against $\mathcal{DP}(p,t)$ of all probes in $p \in P$, and figure (b) depicts a scatterplot of $\overline{R(p)}$ against $\mathcal{DP}(p,t)$ of all probes in $p \in P_g$ such that there exists a probe $p' \in P_g$ with $\mathcal{DP}(p',t') > -22$, for matching tag t'.

able to discriminate). We hypothesize that such a probe will under-represent the true amount of tags present in the sample and will therefore always have a very low rank. The probability vectors depicted in Figure 25 confirm this hypothesis, and confirm the claim that the background distribution of ranks of such a probe p' is affected by the ΔG of the probe-tag pair (p',t') as shown in Section 3.2.

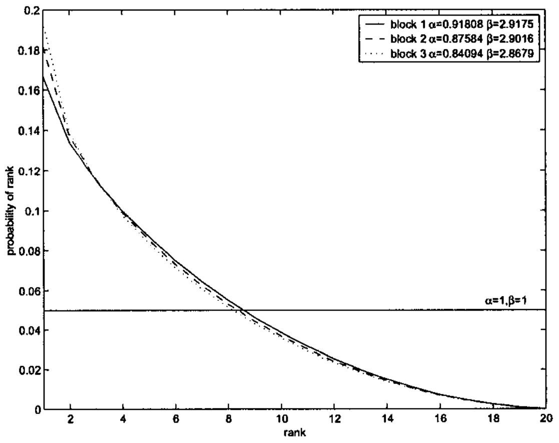


Figure 25: Estimated discretized probability vectors $\phi_{\hat{\alpha},\hat{\beta}}$ for all probes $p \in P$ such that $\mathcal{DP}(p,t) > -22$, for probe-tag pair (p,t).

7 Contribution

There are several technical challenges that arise when dealing with large probe sets and large sets of hybridization experiments due to their size. A significant contribution of this paper is the development of the necessary computational infrastructure to facilitate the testing of data from a large set of hybridization experiments. The chip and hybridization data is stored in a *MySQL* database. We implemented fast algorithms for performing the dynamic programming functions $\mathcal{DP}$ and $\mathcal{S}$ in Java. We also designed and implemented efficient algorithms for the estimation of α and β parameters for a beta distribution from hybridization data. The over 8 billion pairwise operations $\mathcal{X}, \mathcal{D}, \widehat{\mathcal{X}}$ and $\widehat{\mathcal{D}}$ were distributed using BioOpera Process Support for Bioinformatics [3]. The computations were executed on several different machines at the McGill Centre for Bioinformatics Computational Biology Lab. All software was developed using the Apple Project Builder environment on a Macintosh running the Macintosh OS X operating system. We have developed an online resource for the work presented in this thesis and hope to make many of the tools that we created for this framework publicly available [21].

8 Conclusion

We present a framework for detecting degenerate probes in Affymetrix DNA chips. Our predictions are based on a theoretical model of nucleic acid hybridization and on patterns in a large collection of hybridization data. As the number of hybridizations increases, the framework should give increasingly better predictions of degenerate probe-tag pairs. In this sense, we learn from experience the behavior of the chip as it is used increasingly by the community and, in turn, this will enable us to design better future chips.

The method of examining probe ranks to analyze the behaviour of a probe over a set of hybridization data is a simple and successful method. The method was used to define unique patterns of four basic types of degeneracy in terms of distribution of rank. The results from Sections 2 and 3 show a clear relationship between theoretical predictions of degeneracy caused by hybridization and observed ranks. The structural analysis of the conflict graph for HuGeneFL produced several key insights that lead to better prediction strategies. We find that very strong probe-tag pairs (low ΔG) are more frequently predicted to be degenerate than mid-range or weak probe-tag pairs (higher ΔG). We also see that the distribution of ranks of a probe is dependent on the hybridization strength between the probe and the matching tag.

The accuracy of our proposed framework will improve as the number of hybridizations and diversity of conditions under which the hybridizations were performed increases. As described in Section 3, the behaviour of a degenerate probe changes over a range of target intensities. 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). It must be the case that there is sufficient biological diversity in order to be able to properly evaluate the behaviour of a probe over a set of hybridizations. 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 which leads to all non-zero values in the rank count vectors. It is also important to have an abundant set of hybridizations, since as the number of hybridizations increases (thus increasing the expected amount of biological diversity), the likelihood that a well-functioning probe-tag pair always exhibits a degenerate pattern should decrease. Thus, our confidence that these probe-tag pairs are not behaving appropriately is strengthened. The experiments contained in this paper were carried out with a relatively small set of hybridizations (126). A much large collection of hybridizations (on the order of 2000) would ensure that sufficient biological diversity exists so that each probe is expressed (either highly or lowly) at least once.

Though the experimental support functions proved indiscriminate, we believe that the model itself is sound and can lead to more meaningful results through the use of statistically sound methods for determining the appropriate number of blocks and for choosing the theoretical and experimental ΔG thresholds. The most immediate improvement to be made is the discovery of successful experimental support functions. There are several methods we could employ to achieve successful experimental support functions. The most simple approach would be to attain additional hybridization results from experiments performed with HuGeneFL. We may also consider introducing pseudo-counts to increase the amount of data in the rank count vector while preserving the underlying pattern in the vector. There is also the option of employing an alternative test statistic. In designing these experiments, we tried using the 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. Also, the test should distinguish between two hypotheses (that the probe is degenerate or that the probe is non-degenerate) instead of simply testing one hypothesis. We hope to refine our current test statistic and research other non-parametric methods for determining an underlying pattern from an observed rank count vector.

An online resource containing additional results is publicly available at [21] 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. We have made several clarifications and improvements to the framework in [22].

This model can be used for a variety of purposes other than error detection. The versatility of this model was shown in Section 6 when it was used for the analysis of probes that are measured to be indiscriminate by Affymetrix's statistical methods. One can quickly imagine several other useful patterns to detect in the hybridization data. 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 [19].

The focus in [23] is the integration of our framework with the model-based analysis of oligonucleotide arrays from Li and Wong [8, 9]. Li and Wong give a simple model for determining the intensity measurement for a target as a non-linear 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.

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. Ultimately, to validate this model, the candidate degenerate probes and probes pairs must be verified in a wet-lab to conclude whether they are truly degenerate. Based on further results, we could test the validity of this model by creating a chip with known degenerate probes or probe pairs and run a variety of hybridizations to guarantee biological diversity.

A Nucleic Acid Sequence Hybridization

The technology of microarrays is predicated upon the *hybridization* of two single stranded nucleotide sequences into stable double stranded complexes. Such a duplex is created by the formation of hydrogen bonds between the Watson-Crick base pairs of two strands of DNA, two strands of RNA or one strand each of DNA and RNA. The hybridization reaction takes place under specific physical and chemical conditions that can be altered to promote or inhibit the formation of the duplex.

The *Watson-Crick (wc-) complement* $\bar{s}$ of a nucleic acid sequence s is obtained by reversing the sequence and interchanging $A \leftrightarrow T$ (or U in the case of RNA) and $C \leftrightarrow G$.

Example: If $s = ACTTGCAATCGTAATCGC$ then $\bar{s} = t = GCGATTACGATTGCAAGT$

A **Watson-Crick complement pair** of oligonucleotides s and t, $s = \bar{t}$ may exhibit **perfect hybridization** to form a stable duplex. For example, $s = ACTTGCAATCGTAATCGC$, $t = GCGATTACGATTGCAAGT$ hybridize as follows:

```
5' - A-C-T-T-G-C-A-A-T-C-G-T-A-A-T-C-G-C - 3'
     | | | | | | | | | | | | | | | | | |
3' - T-G-A-A-C-G-T-T-A-G-C-A-T-T-A-G-C-G - 5'
```

Two oligonucleotides s and t may hybridize even if not all of their aligned bases are Watson-Crick pairs. However, imperfect hybridization or hybridization with mismatched base pairs form less stable duplexes.

B Thermodynamic Models of Nucleic Acid Hybridization

Nucleic acid sequence hybridization and DNA secondary structure formation are well-understood processes. There exist several models for predicting either the stability of a candidate structure. [10, 13, 15, 17, 24, 28].

A **combinatorial model** for DNA hybridization behaviour simply counts the number of base pairs that form Watson-Crick pairs and those that mismatch. Such a model is not based on the biochemical properties that govern the hybridizations, but is rather based on the similarity of the two tags in question. Such a model commonly uses measurements of distance between strings to predict the occurrence of imperfect hybridizations (Hamming distance, longest common substring or subsequence). We have seen that two very similar targets will cross-hybridize to each other's probes. It is very likely that such targets will share a long common substring, though there may be a possibility for cross-hybridization when there are many shorter common substrings and weakly mismatched base pairs.

Several models have been proposed in the literature to analyze DNA hybridization by estimating thermodynamic properties governing the formation of a duplex. The **melting temperature** T_M of two strands of nucleic acid is the temperature at which half of each strand's nucleotides are in double-helical duplex form and half are in single-stranded form. The most simple model for estimating the T_M of a duplex is the **2-4 model** consisting of a simple rule based on the number of hydrogen bonds between a two nucleotides. With the **2-4 model**, the T_M of a sequence and its Watson-Crick complement is $T_M = 2(\text{number of A-T base pairs}) + 4(\text{number of C-G base pairs})$ [24].

A thermodynamic **nearest-neighbour model** of DNA hybridization offers a more specific and biologically based model for short DNA sequences. The nearest-neighbour model is widely applied for estimating thermodynamic properties of duplex formation. The model can be used to measure the stability of duplex formation between two short stands of nucleic acid and can also be used for estimating the stability of a folded secondary structure. The nearest-neighbour model can be used to determine several thermodynamic measures of helix formation and helix melting ($\Delta S^\circ, \Delta H^\circ, \Delta G^\circ$ and T_M). We use change in free-energy ΔG (measured in *kcal/mol*) of the helix formation to provide a relative measure of duplex stability. The change in free-energy is more accurate than other thermodynamic measures [15]. The nearest-neighbour model defines ΔG°_{37} as the sum of the following terms: a helix initiation penalty, a sum of free-energy change for helix formation at each base pair, a penalty for symmetric strands and if applicable, a penalty for the terminal dangling ends or external mismatches. The free-energy charge for each base pair is determined by summing ΔG nearest-neighbour parameters calculated for helix formation at 37°.

DNA and RNA helices have different structures and accordingly, have different nearest-neighbour parameters. They also differ in helix initiation factors. Several sets of nearest neighbour thermodynamic parameters have been obtained for DNA/DNA, RNA/RNA and DNA/RNA helices. All three types of duplexes are possible in the context of microarrays. With Affymetrix GeneChips, there is the possibility for RNA/DNA and RNA/RNA hybridizations. RNA/RNA duplexes are generally most stable, and the stability differences between DNA/DNA and RNA/DNA depend on their sequence composition [25]. We use a set of DNA/DNA parameters determined at John SantaLucia's lab because in addition to Watson-Crick pairs, they comprise mismatched pairs and dangling end parameters [14].

C Affymetrix GeneChips

Affymetrix GeneChips are the most commercially popular DNA chips for gene expression assays. The technology allows expression states of genes to be determined by measuring the amount of expressed mRNA captured by the chip. The hybridization occurs between the cRNA strand and the immobilized DNA probe. The GeneChip experimental process proceeds as follows. A sufficiently large sample of mRNA transcripts from the target tissue or organism is processed and transformed into a target sample of labeled cRNA fragments. This target sample is then brought into contact with the chip and the transcript fragments theoretically hybridize with (and only with) their wc-complementary probes on the chip. The hybridization that occurs on the chip forms RNA/DNA hybrids. After hybridization occurs and the chip has been stained, the intensity of each probe location is optically measured. The intensity readings of each set of probes representing one expression target are statistically analyzed to classify the transcript as present or absent, and to estimate that gene's quantitative level of expression [12].

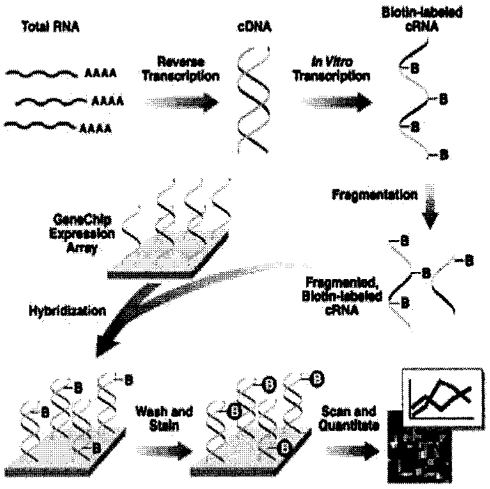


Figure 26: The process of an Affymetrix GeneChip expression experiment. Image from [12].

References

[1] Affymetrix White Paper: Statistical Algorithms Description Document. (2002) *Affymetrix, Inc.*

[2] BenDor, A. et al. (2000) Universal tag-antitag systems: A combinatorial design scheme. *J. Comput. Biol.* 7:503-519.

[3] BioOpera website `http://www.inf.ethz.ch/personal/bausch/bioopera/main.html`, ETH Department of Computer Science.

[4] Personal communication between Michael Hallett, Tom Hudson and Robert Sladek (2003).

[5] Hastings, N., Peacock, J. (1975) Statistical distributions: a handbook for students and practitioners. *Halsted Press, New York, U.S.A., 1975.*

[6] Hubbell, E., Pevzner, P. (1999) Fidelity probes for DNA arrays. *Proc. 7th International Conference on Intelligent Systems for Molecular Biology.* AAAI Pres. Heidelberg, Germany, August, 113-117.

[7] Lemon, W. J., Palatini, J. J. T., et al. (2001) Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays. *Unpublished*

[8] Li, C. and Wong, W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS*, 98, 1, 31-36.

[9] Li, C. and Wong, W. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2, 8, 1-11.

[10] Mathews, D. H., Sabina, J., Zuker, M., Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288, 910-940.

[11] Moore, S. et al. (1995) The basic practice of statistics. *W. H. Freeman and Company, New York.*

[12] NetAffx website (2002) `http://www.netaffx.com/`, Affymetrix Inc. 2001

[13] Sankoff, D., Zuker, M. (1984) RNA Secondary structures and their prediction. *Bull. Mathematical Biology* 46, 591-621.

[14] SantaLucia, J.Jr., Allawi, H., Seneviratne, A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability *American Chemical Society Biochemistry* 96, 35, 3555-3562.

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

[16] SantaLucia, J.Jr. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with Watson-Crick base paring. *American Chemical Society Biochemistry*, 37, 14719-14735.

[17] SantaLucia, J. Jr. (2002) Loop parameters. *Unpublished data.* `http://bioinfo.math.rpi.edu/z̃ukerm/dna/credit.html`

[18] Sengupta, R., 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.

[19] Shen-Orr, S., Milo, R., Mangan, S., Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* 31:64-68 (2002)

[20] Ship, M, Ross, K, et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene expression Profiling and Supervised Machine Learning *Nature Medicine* 8, 1, January 2002.

[21] Smith, K., and Hallett, M. (2003) Online resource for paper `http://www.mcb.mcgill.ca/~genechips`, McGill University.

[22] Smith, K., and Hallett, M. (2003) Towards quality control in DNA Microarrays. *Submitted to Journal of Computational Biology.*

[23] Smith, K. and Hallett, M. (2003) Estimating probe parameters for intensity computations. *In preparation.*

[24] Strachen, T., Read, A. (1996) Human molecular genetics. *Bios Scientific Publishers*

[25] Sugimoto, N., Nakano, S., et al. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *American Chemical Society Biochemistry* 34, 11211-11216.

[26] Tobler, J. B., Molla, M. N., Nuwaysir, E. F., Green, R. D., Shavlik, J. .W. (2002) Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *ISMB 2002*, August, Edmonton, pp. 164–171.

[27] Virtaneva, K. et al. (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.

[28] Zuker, M., Mathews, D. H., Turner, D. H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: *RNA biochemistry and biotechnology*, J. Barciszewski and B. F. C. Clark, eds. NATO ASI Series, Kluwer Academic Publishers.