

# Designing Specific Oligonucleotide Probes for the Entire *S. cerevisiae* Transcriptome

Doron Lipson<sup>1</sup>, Peter Webb<sup>2</sup>, and Zohar Yakhini<sup>1,2</sup>

<sup>1</sup> Technion, Haifa 32000, Israel

dlipson@cs.technion.ac.il

<sup>2</sup> Agilent Laboratories

{peter\_webb, zohar\_yakhini}@agilent.com

**Abstract.** Probe specificity plays a central role in designing accurate microarray hybridization assays. Current literature on specific probe design studies algorithmic approaches and their relationship with hybridization thermodynamics. In this work we address probe specificity properties under a stochastic model assumption and compare the results to actual behavior in genomic data. We develop efficient specificity search algorithms. Our methods incorporate existing transcript expression level data and handle a variety of cross-hybridization models. We analyze the performance of our methods. Applying our algorithm to the entire *S. cerevisiae* transcriptome we provide probe specificity maps for all yeast ORFs that may be used as the basis for selection of sensitive probes.

## 1 Introduction

A *probe*, in the context of the current study is a (nucleic acid) molecule that strongly interacts with a specific target in a detectable and quantifiable manner. Oligonucleotides are used as probes in an increasing number of molecular biology techniques. They are central in Southern and northern blotting, in *in-situ* hybridization assays, in quantitative PCR, and in array based hybridization assays (chips) where they are immobilized on a surface [1, 2]. Some applications and protocols require probe labeling while in others (e.g. arrays) the target is being labeled.

There are, roughly speaking, two parameters by which to evaluate candidate probes for a given application. *Sensitivity* - is it really going to strongly interact with its target, under the assay's conditions, and how much target is needed for the reaction to be detectable or quantifiable; and *specificity* - how well does the probe discriminate between its intended target and other messages that it might cross hybridize to. This study addresses specificity design questions.

A particular application in which design and specificity issues arise is gene expression profiling, in which the particular subset of genes expressed at a given stage and its quantitative composition are queried. Such information can help in characterizing sequence to function relationships, in determining effects (and side effects) of experimental treatments, and in understanding other molecular

biological processes [3, 4], many with clinical implications [5, 6]. Gene expression profiling is typically performed using array based hybridization assays. The actual probes that populate a custom designed expression profiling array are specifically designed and chosen to measure the expression levels of a defined set of genes. Given the state of the human genome sequence draft and the extent of gene hunting efforts currently invested by the scientific community, it is reasonable to approach gene expression profiling assuming complete knowledge of the sequences of the genes of interest, as well as those of many of the genes expressed in the background message.

Let  $\mathcal{N} = A, C, G, T$  be the alphabet representing the four different nucleotides. Our general design question is as follows. We are given a *target gene*,  $g$  - a sequence over  $\mathcal{N}$ , of length 500-10000; and a *background message*,  $D$  - a large set of sequences (with lengths in the same range), representing all possible mRNA molecules that might be active in our sample<sup>3</sup>. In  $g$  we seek substrings that represent Watson-Crick (WC) complements to molecules that have a high WC mismatch to the background message. These substrings are presumably good probe binding site candidates, in terms of specificity. Since they don't have a close WC match in the background message, they do not have a high cross-hybridization potential. An equivalent computational task is to find, in  $g$ , the substrings that are far, in Hamming distance, from the background message (as a set). We seek many symbolically specific probe candidates since specificity screening is only one stage in the probe selection process. Specificity issues arise in probe design as described above (and in other forms), as well as in the design of PCR primers [7] and the design and development of anti-sense drugs.

We re-emphasize that distance (as above, in the Watson-Crick or Hamming sense) is not the only parameter that determines the specificity of a probe candidate. The issues of multiplicity (how many times do mismatches occur in the background), abundance (how strongly expressed are genes with close matches), hybridization potential (how competitive is the background, in terms of thermodynamics and not only homology), and others, play an important if not well understood and not easily quantified role.

Previous work [8] presents an algorithm for selecting specific probes using suffix trees and a model of hybridization thermodynamic. The authors apply the algorithm on several genomes providing few candidate probes per gene. Other methods are described in [9]. In this study we address additional issues:

- A process for specific probe design, whose output is specificity map of all possible probes for any specific gene, based on Hamming distance from the background message or other possible thermodynamic models. This map can be used as a basis for selecting sensitive probes.
- Analysis of the efficiency and effectiveness of the proposed algorithm.
- Consideration of the relative abundance of background transcripts by applying different filters to different relative expression ratios. Abundance, for this purpose, is obtained from pre-existing data. Such data is expected to grow

---

<sup>3</sup> Typically, we shall use some database that stores much of (what is known of) this organism/tissue/stage specific information

as more expression measurements are performed and as databases storing this information continue to evolve.

- Rigorous stochastic analysis of the statistical protection provided by probe length when pre-existing background sequence data is not sufficient.

In Section 2 we study a stochastic model that indicates how long we should expect probes with a given threshold specificity to be. In Section 3 we present algorithmic approaches to assessing probe candidate specificity against available data. We analyze some efficient and effective heuristics and modifications that address transcript abundance and thermodynamic models. Finally, in Section 4, we discuss implementations and the design of specific probes for the entire *S. cerevisiae* transcriptome.

## 2 Statistical Properties of Symbolic Specificity

In this section we analyze the expected behavior of a probe candidate mismatch to the background message, under uniform stochastic model assumptions. Understanding this behavior is instrumental in determining practical probe lengths that provide statistical defense against cross hybridization. Probe length decisions are central to any array development and manufacturing process. Probe lengths influence pricing, probe fidelity<sup>4</sup> and other parameters. Clearly - the uniform distribution assumed in the model is certainly not the observed reality. We therefore test the quality of the stochastic model predictions and their validity for genomic data.

We denote the Hamming distance between two strings  $s, t \in \mathcal{N}^k$  by  $H_k(s, t)$ . More precisely, if  $s = s_1 \dots s_k$  and  $t = t_1 \dots t_k$  then  $H_k(s, t) = \sum_{i=1}^k 1_{[s_i \neq t_i]}$ .

When considering a probe candidate and a given background message, we are interested in the distance between the probe and the entire message. For a string  $s \in \mathcal{N}^k$  and a string  $B \in \mathcal{N}^N$ , where  $\mathcal{N} \gg k$ , we set

$$d(s, B) = \min_{1 \leq i \leq N-k+1} H_k(s, B_i^{i+k-1}).$$

This is analogous to treating  $B$  as a set of  $k$ -long strings and measuring the Hamming distance from  $s$  to the said set.

In our stochastic model we assume that both  $g$ , the gene of interest, and  $B$ , the background message, were drawn uniformly and independently over  $\mathcal{N}^m$  and  $\mathcal{N}^N$ , respectively. Given  $k$  and  $N$  as set forth consider the random variable  $d_{N,k} = d(p, B)$ , where  $p$  is a probe candidate, a  $k$ -substring of  $g$ . The distribution of  $d_{N,k}$  gives us starting points for answering the probe design and length determination questions described in the introduction.

### 2.1 Poisson Approximation

Our study of the distribution of  $d_{N,k}$  utilizes approximation methods that are part of a general approximation theory developed by Stein and others [10, 11]. For completeness we state the results that are directly applicable in our context.

<sup>4</sup> Fraction of full-length probe in any nominal feature.

Let  $\Gamma$  be an index set. Let  $I_i : i \in \Gamma$  be a collection of indicator random variables on some probability space. Let  $W = \sum_{i \in \Gamma} I_i$ . Let  $\lambda = E(W)$ . Assume that for every  $i$  there is a set of indices  $\Gamma_i$  such that  $I_i$  is independent of  $I_j : j \notin \Gamma_i$ .  $\Gamma_i$  is called  $I_i$ 's *dependence neighborhood*.

**Theorem:** Let  $W$  as above. Let  $Z \sim \text{Poisson}(\lambda)$ .

Set  $b_1 = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} E(I_i)E(I_j)$  and  $b_2 = \sum_{i \in \Gamma} E(I_i) \sum_{j \in \Gamma_i} E(I_j | I_i = 1)$ .

Then  $|\text{Pr}(W = 0) - e^{-\lambda}| \leq (1 \wedge \lambda^{-1})(b_1 + b_2)$ .

*Remark:* The independence assumption can, indeed, be relaxed. An additional error term is then added.

For  $r = 0, 1, 2 \dots k$ , let  $\mu_r = \frac{1}{4^k} \sum_{\rho=0}^r \binom{k}{\rho} \cdot 3^\rho$  and  $\lambda(N, r) = (N - k + 1)\mu_r$ .

**Theorem:** Fix a word  $s_0$  of length  $k$ . Consider the random variable  $d_N(s_0)$  as above. For  $r = 0, 1, 2, \dots, k$ , let  $F(r) = \text{Pr}(d_N \leq r)$ .

Set  $b_1(r) = (N - k + 1)(2k - 1)\mu_r^2$ .

Then  $|F(r) - (1 - e^{-\lambda(N, r)})| \leq (1 \wedge \lambda(N, r)^{-1})b_1(r)$ .

**Proof:** For  $r = 1, 2, \dots, k$  define random variables

$$V_N(r) = \sum_{i=1}^{N-k+1} 1_{B_r(s_0)}(B_i^{i+k-1}),$$

Where  $B_r(s_0)$  is the Hamming ball of radius  $r$  around  $s_0$ , in  $\mathcal{N}^k$ . We then have  $1 - F(r) = \text{Pr}(V_N(r) = 0)$ .  $V_N(r)$  is a sum of indicator random variables. Clearly  $E(V_N(r)) = \lambda(N, r)$ . A Poisson distribution is not always a good approximation for that of  $V_N(r)$ . This can be seen by considering the case  $r = 0$  and a word  $s_0$  with small periods. In this case  $V_N(r)$  will have compound Poisson distribution. However, we are only interested in estimating the mass at 0.

For  $1 \leq i \leq N - k + 1$  set

$$J_i = 1_{B_r(s_0)}(B_i^{i+k-1}) \cdot \left( \prod_{j=i+1}^{i+k-1} (1 - (1_{B_r(s_0)}(B_j^{j+k-1}))) \right) \text{ and } U_N(r) = \sum_{i=1}^{N-k+1} J_i$$

$U_N(r)$  counts the left-most visits to the Hamming ball around  $s_0$ , encountered in the background sequence  $B$ . We will assess a Poisson approximation to  $U_N(r)$ . Since  $\text{Pr}(V_N(r) = 0) = \text{Pr}(U_N(r) = 0)$  we will have an estimate of the quantity of interest.

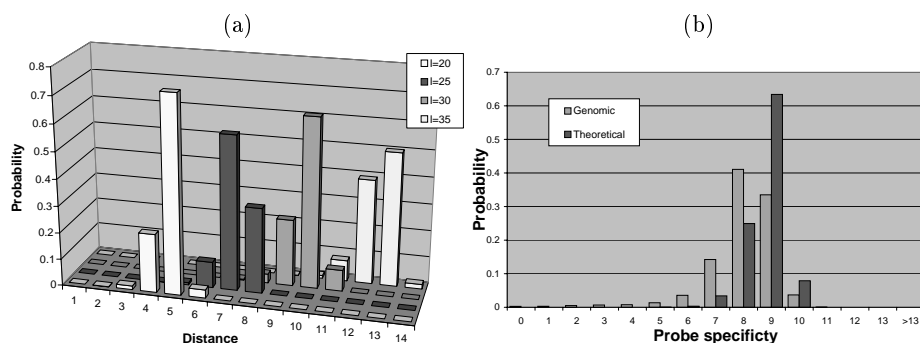
To apply the Poisson approximation theorem stated above we use the dependence neighborhoods  $\Gamma_i = \{1 \leq j \leq N = k + 1 : |j - i| \leq k\}$ .

Estimating the bounds we get  $b_1 = (N - k + 1)(2k - 1)\mu_r^2$ .

By stationarity:  $b_2 = 2(N - k + 1)\mu_r \sum_{j=2}^k E(I_j | I_1 = 1)$ .

The summands here are all 0 since the presence of  $s_0$  at a given position means a left-most visit can not occur in the next  $k - 1$  positions. Thus  $b_2 = 0$ .

The above approximation result enables us to compute the distribution of  $d_{N,k}$  for various assumed transcriptome sizes and various probe lengths, as shown in Figure 1a. In Figure 1b we compare the results obtained for the stochastic model to these actually observed when measuring distances of candidate probes to the rest of the yeast transcriptome. The reasons for the observed bias are discussed in the next section. Despite the bias, it is clear that the model calculation provides us with information on the length needed for desired statistical protection against cross hybridization. Model results should be taken as over-estimates of the actual distances.



**Fig. 1.** a) Expected distribution of probe specificities (given in distance from background message of length  $N=9\text{Mb}$  - the size of the yeast transcriptome), for probes of different lengths  $l$ . b) Distribution of probe specificity according to theoretical computation in comparison to the specificity distribution retrieved by a search over a genomic database (150 ORFs of *S. cerevisiae*)

### 3 Algorithmic Approach to Evaluating the True Symbolic Specificity

In the previous section, we analyzed the statistical properties of an unknown background, using a stochastic model. However, as can be seen from the comparison to actual biological data (Figure 1b), the model predicted probe specificity distribution deviates from the true distribution. A remarkable difference is the noticeable fraction of low specificity probes in the genomic distribution. These are virtually probability 0 events, in the model. The simple and straightforward explanation for this difference is that the genomic data is structured and therefore is expected to contain more similarities than random data. Specifically, since genomic sequences are the results of an evolutionary process, total and partial

duplications of genes are not uncommon, resulting in repetitions that are very improbable, under our stochastic model (in particular, probes with 0 specificity).

However, as an increasing number of genomes are fully sequenced and annotated, the true background message for these organisms can be evaluated by means of a search algorithm. As mentioned in Section 1, we will use Hamming distance as the proxy parameter defining probe specificity. The specificity search attempts to compute the Hamming distance of all probe-candidates of a given length to the background message (assumed given in some accessible data structure). The results of such a computation will directly entail the identification of good (specific, in Hamming terms) probes. Providing the length of probe candidates was set according to the results of Section 2, then we expect plenty of good probes. Later in this section we will also address the issue of gene abundance as a factor influencing probe specificity.

### 3.1 Exhaustive Search Algorithm

We now state the computational problem in more formal terms. We address the case in which the set of candidate probes is the full set of substrings of  $g$ , of a given length.

**Full Distance Search** (Statement of the problem)

**Input:** The target gene  $g$  (of length  $m$ ), a background message  $D$  (of length  $N$ )<sup>5</sup>, a probe length  $l$ .

**Output:** A set of Hamming distances  $\{d(g_i^{i+l-1}, D) : 1 \leq i \leq m - l + 1\}$ .

In practice,  $N \approx 1 - 100Mb$  and  $m \approx 1 - 10Kb$ .

The naive algorithm performs an exhaustive search of all candidate probes against the background message, computing all Hamming distances as above.

### 3.2 Probe Classification Algorithm

Recall, that we want to design specific probes for  $g$ . This means that we are not really interested in computing the set of exact distances  $\{d(g_i^{i+l-1}, D) : 1 \leq i \leq m - l + 1\}$ , but only in classifying the probe candidates to those of small distance (bad probes) and those of large distance (good probes). Formally, the computational problem reduces to:

**Classification of Probe Candidates** (Statement of the problem)

**Input:** The target gene  $g$  (of length  $m$ ), a background message  $D$  (of length  $N$ ), a probe length  $l$ , a quality threshold  $r$ .

**Output:** A partition of the set  $\{1 \leq i \leq m - l + 1\}$  into  $B = \{i : d(g_i^{i+l-1}, D) < r\}$  and  $G = \{i : d(g_i^{i+l-1}, D) \geq r\}$

This formulation of the problem brings about a simple but useful observation:

---

<sup>5</sup>  $D$  may be given as a set of sequences with a total length  $N$ , or as a concatenated sequence

*Observation:* If two strings have  $H_l(s_1^l, t_1^l) \leq d$ , then there must be at least one substring of  $s$ , of length  $\geq \lceil (l-d)/(d+1) \rceil$ , which is a perfect match to the corresponding substring of  $t$ . This observation comes from the fact that the worst distribution of  $d$  mismatches (i.e. the distribution that produces the shortest maximal perfect match) is the one in which the mismatches are distributed evenly along the string - between intervals of matches that are  $\lceil (l-d)/(d+1) \rceil$  long.

Therefore, to classify probes we can perform an indexed search. A version that is applicable here is formally described in **IClaS** (Indexed Classified Search - Figure 2). Here is an informal description: Generate and index of all words of a given length, to be called the *seed length*, and denoted  $\xi$ . In each entry of this index keep a list of the occurrences of the corresponding word in the target gene  $g$ . Scan all the background message with a window of size  $\xi$ . For each such location, go to all the occurrences in  $g$  of the encountered word, grow  $l$ -strings around the two short words (in  $g$  and in  $D$ ) in parallel, compute the distances and update the distance vector when necessary (that is, whenever the currently interrogated  $l$ -strings have a match closer than anything previously encountered).

Let $I$ be an integer lists, with $4^\xi$ entries, addressable as words in $\mathcal{N}^\xi$ .	
<p><i>Pre-Processing</i></p> <pre>for (i=0; i&lt;m-ξ+1; i++)     Insert(i, I(g_i^{i+ξ+1}))</pre> <p><i>Initializing d</i></p> <pre>for (i=0; i&lt;m-ξ+1; i++)     d(i)=1</pre>	<p><i>Scanning</i></p> <pre>for (j=0; j&lt;N-ξ+1; j++)     for (i∈I(D_j^{j+ξ-1}))         ld=d(g_i^{j+l-1}, D_j^{j+l-1})         for (a=0; a&lt;l-ξ; a++)             d(i-a)=min(d(i-a), ld)             ld=ld-1_{(g_{i+l-1-a}≠D_{j+l-1-a})} + 1_{(g_{i+l-a}≠D_{j+l-a})}</pre>

**Fig. 2. IClaS:** Computing some (but possibly not all) distances and classifying probe candidates, using an indexed search

**Performance** The performance of **IClaS** can be analyzed by counting the number of probe-length comparisons performed by the algorithm. Clearly, the exhaustive search performs  $O(mN)$  such comparisons<sup>6</sup>. The number of comparisons made by **IClaS** cannot be determined directly, since it depends on the actual input and more specifically on the total number of index entries each one of the seed  $\xi$ -mers. We analyze the order of the expected number of such comparisons assuming that  $g$  and  $D$  are uniformly and independently drawn: Given a seed  $s$ , set  $\gamma(s)$  to be the number of occurrences of  $s$  in  $g$  (i.e. the number of index entries for  $s$ ). Set  $\delta(s)$  to be the number of occurrences of  $s$  in  $D$ . For this  $s$ , for each of

<sup>6</sup> The performance of both **IClaS** and the exhaustive search can be improved slightly by using various implementation "shortcuts". These do not affect the running time by more than a constant factor.

its appearances in  $D$ ,  $l - \xi + 1$  comparisons are performed for each index entry of  $s$ . The total number of probe-length comparisons performed by the algorithm is therefore in the order of  $\sum_{s \in N^\xi} l \cdot \gamma(s) \cdot \delta(s)$ . Employing our uniformity and independence assumptions we calculate the expected number of such comparisons to be:  $E\left(\sum_{s \in N^\xi} l \cdot \gamma(s) \cdot \delta(s)\right) = l \cdot \sum_{s \in N^\xi} E(\gamma(s)) \cdot E(\delta(s)) = l \cdot 4^\xi \frac{m}{4^\xi} \frac{N}{4^\xi} = \frac{lmN}{4^\xi}$

Thus, the relative efficiency of **IClaS** over the exhaustive search may be expected to be in the order of  $l/4^\xi$ , which becomes very significant as  $\xi$  grows.

Testing of the actual performance of **IClaS** for various values of  $\xi$ , in comparison to the actual performance of the exhaustive search, was performed on a Pentium III 650MHz. For an arbitrary *S. cerevisiae* transcript of length  $m = 300$  against the entire transcriptome with  $l = 30$  the following running times (in seconds) were measured for  $\xi = 9, 8, 7, 6, 5, 4$  and an exhaustive search, respectively: 73, 79, 116, 283, 781, 2750 (46 mins), and 23011 (6.4 hrs).

**Reliability** In the previous section we observed that working with a small seed size,  $\xi$ , will guarantee catching all bad (not specific enough) probes in a target gene. In the previous section we saw, however, that the time complexity of the indexed search strongly depends on the size of the seed. In this section we examine the connection between  $\xi$  and the reliability of the corresponding **IClaS** indexed search. The reliability is measured in terms of the probability of false positives: bad probes that were not caught by **IClaS**.

We start by stating a combinatorial result, a special case of the problems studied in [12, 13]:

*Definition:* Given a string  $s$  over an alphabet  $\Sigma$  and  $\sigma \in \Sigma$ , a substring  $s_i^j$  is a *leftmost run* of  $\sigma$ 's if  $s_i^j \in \{\sigma\}^*$  and either  $s_{i-1} \neq \sigma$  or  $i = 1$ . For example, the string *AABBBBBBBBAA* contains only one leftmost run of 3 *B*'s while the string *AABBBABBBAA* contains two leftmost runs of 3 *B*'s.

Consider all words of length  $l$  over the binary alphabet  $\{A, B\}$ . Consider all such words that have exactly  $l(A)$  occurrences of *A*, and exactly  $l(B) = l - l(A)$  occurrences of *B*. The number of such words that have at least  $j$  leftmost runs of at least  $\sigma$  *B*'s can be computed by taking all words of  $l(A)$  *A*'s and  $(l(B) - \sigma j)$  *B*'s, and for each such word taking all combinations of insertions of the  $j$  runs between the  $l(A)$  *A*'s. In total there are  $\binom{l(A)+1}{j} \binom{l-\sigma j}{l(A)}$  such words.

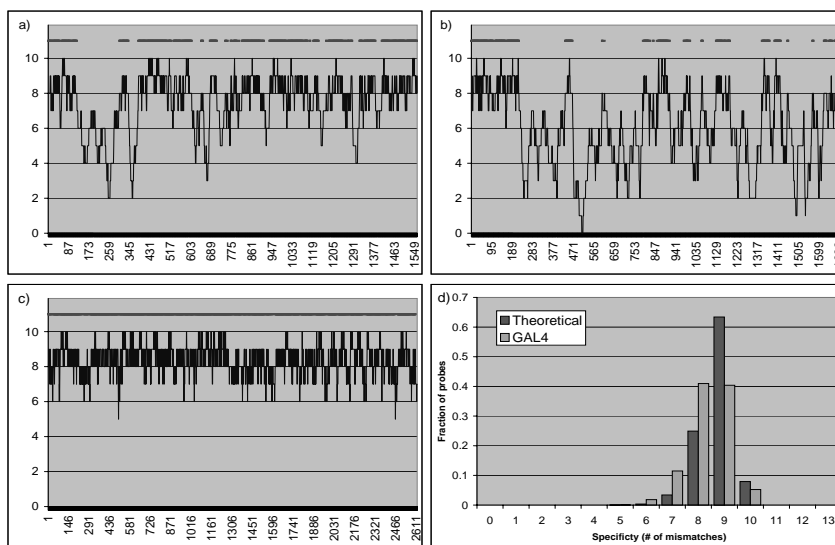
However, the same word may be overcounted in the above expression. For example, for  $j = 1$ , a sequence containing exactly two leftmost runs of at least  $\sigma$  *B*'s will be counted twice - once for containing at least each of the two runs. Using the inclusion and exclusion principle, the number of such words that have at least one run of at least  $\sigma$  *B*'s is:  $\sum_{j \geq 1} (-1)^{j+1} \binom{l(A)+1}{j} \binom{l-\sigma j}{l(A)}$ .

Consequently, for a probe  $p$  of length  $l$  and a seed length  $\xi$ , the probability that a background sequence  $b$  of length  $l$  uniformly drawn from the Hamming sphere of radius  $d$  from  $p$  (i.e.  $H(p, b) = d$ ) will contain a  $\xi$  long match is:

$$\alpha(l, d, \xi) = \frac{\sum_{j \geq 1} (-1)^{j+1} \binom{d+1}{j} \binom{l-\xi j}{d}}{\binom{l}{d}},$$

which represents a lower bound for the reliability of the indexed search.

Given a probe design problem, it is therefore possible to choose a maximal seed length  $\xi$  that ensures the desired reliability  $\alpha$ . As we already noted, biological data does not usually conform to statistical results derived over a uniform distribution of sequences. However, since in this analysis the worst case scenario is the one in which the entropy is maximal, structured biological data may be expected to give better results (higher reliability). Moreover, as indicated in Table 1, even when choosing a seed for  $\alpha = 95\%$ , the reliability of detecting a bad probe is virtually 100%.



**Fig. 3.** Probe distance maps for three genes of the GAL family. For each gene, the specificity of all possible probes (in Hamming distance from the entire transcriptome) are plotted against their position along the sequence. Areas of specific probes (above a threshold  $r = 7$ ) are marked above: (a) the sequence of GAL1 contains an area where probes are less specific (this is an area of homology to GAL3), (b) the sequence of GAL2 contains only small stretches of specific probes (because of high similarity to genes from the HXT family), (c) the sequence of GAL4 is almost totally specific, with a probe specificity distribution similar to theory, as shown in (d).

### 3.3 Addressing Gene Abundance

Choosing the mismatch threshold  $r$  (or equivalently, the seed parameter  $\xi$ ) depends on the degree of interference that may be expected due to cross-hybridization. Gene abundance is therefore significant in evaluating candidate probes. When comparing probes of a gene  $g$  to a background gene  $b$ , compare the situation

**Table 1.** Reliability of **IClaS** algorithm for a search over 7 ORFs selected randomly from the *S. cerevisiae* transcriptome (a total of 12184 probes). Parameters for the search were  $l = 30$ ,  $\alpha = 95\%$ ,  $r = 7/5/3$  ( $e$  denoting the respective seed length), or an exhaustive search. Value in table denotes the cumulative fraction of probes for a given distance  $d$  or less, found in each search. Since the algorithm never overestimates a probe's specificity, results show that the actual reliability is 100% at the desired threshold (emphasized cells).

$d$	Exhaustive	$r = 7, e = 5$	$r = 5, e = 7$	$r = 3, e = 9$
0	<b>0.0034</b>	<b>0.0034</b>	<b>0.0034</b>	<b>0.0034</b>
1	<b>0.0095</b>	<b>0.0095</b>	<b>0.0095</b>	<b>0.0095</b>
2	<b>0.0166</b>	<b>0.0166</b>	<b>0.0166</b>	<b>0.0166</b>
3	<b>0.0208</b>	<b>0.0208</b>	<b>0.0208</b>	<b>0.0208</b>
4	<b>0.0258</b>	<b>0.0258</b>	<b>0.0258</b>	0.0257
5	<b>0.0335</b>	<b>0.0335</b>	<b>0.0335</b>	0.0334
6	<b>0.0543</b>	<b>0.0543</b>	0.0541	0.0502
7	<b>0.1857</b>	<b>0.1857</b>	0.1788	0.1399
8	<b>0.6148</b>	0.6147	0.5767	0.4272
9	<b>0.9592</b>	0.9590	0.9358	0.8285
10	<b>0.9995</b>	0.9995	0.9988	0.9880
11	1	1	1	0.9999

in which  $g$  is of high abundance whilst  $b$  is rare, to the opposite situation in which  $g$  is rare whilst  $b$  is abundant. In the first scenario, cross hybridization between  $b$  and the probe designed for  $g$  is unlikely to significantly interfere with the correct hybridization between the probe and  $g$  itself. Accordingly, a choice of a low mismatch threshold  $r$  (or a long seed length  $\xi$ ) is appropriate for this situation. In the second scenario  $b$  is fierce competition for  $g$  over the probe and a high mismatch threshold  $r$  (a short seed length  $\xi$ ) is required.

Since gene expression profiling experiments are specifically intended to measure the abundance of gene transcripts it is clearly impossible to assume full a-priori knowledge of this information. However, for some organisms (e.g. *S. cerevisiae*) there exists substantial experimental information about the typical abundance of gene transcripts. In such cases, rough relative abundances of the different genes may be used for a wiser choice of parameters  $r$  and  $\xi$ . In addition, abundance information may be used in an iterative design process. Fortunately, **IClaS** easily lends itself to such modification. An important observation is that although gene abundance may be regarded as a continuous parameter, the parameters  $r$  and  $\xi$  themselves are discrete, and usually fall within a short range (e.g. for  $l = 30$   $r$  is usually confined to the range 4-8). Therefore, rather than use some continuous function that calculates the optimal  $r$  for the abundance levels of  $g$  and each  $b \in D$ , it is equivalent and simpler to make use of a direct lookup table for determining  $r$ .

**Classification of Probe Candidates by Abundance Weighting** (Statement of the problem)

**Input:** The target gene  $g$  (of length  $m$ ), the background message  $D$  partitioned into  $\{D_1, D_2, \dots, D_k\}$  where each  $D_i$  is an abundance category, a probe length  $l$ , a vector  $(r_1, r_2, \dots, r_k)$  where  $r_i$  is the quality threshold for abundance group  $D_i$ .

**Output:** A partition of the set  $\{1 \leq i \leq m-l+1\}$  into  $B = \{i : \exists j d(g_i^{i+l-1}, D_j) < r_j\}$  and  $G = \{i : \forall j d(g_i^{i+l-1}, D_j) < r_j\}$ .

The algorithm **IClaSA** is a loop of runs of **IClaS**, where in each iteration the complete set of probes of  $g$  is compared to  $D_i$  using the threshold  $r_i$  to produce the appropriate vector of approximate distances.

### 3.4 Classifying Probes by Melting Temperature of Potential Cross-hybridization Nucleation Complexes

As shown, the algorithm for classifying probes according to a mismatch threshold  $r$  is analogous to classifying the same probes according to a seed length  $\xi$  that may be calculated from  $r$ . For many practical applications  $\xi$  is even more appropriate as a classification parameter since cross hybridization to a probe (of a non-specific message) is most probably the result of hybridization initiated around a perfect match nucleus (of length  $\xi$ )<sup>7</sup>. In this approach we directly set the seed parameter  $\xi$  than calculate it using  $r$ . A different, perhaps more realistic, method of classifying probes is according to the melting temperature ( $T_m$ ) of a nucleation complex that may initiate cross-hybridization. Following this reason, we define the proximity of a probe  $p$  from the background  $B$  to be the maximal melting temperature of a subsequence of  $p$  that is a perfect match to some background subsequence. Formally:

$$f(p, D) = \max_{i,k} \left\{ T_m(p_i^{i+k-1}) : \exists d \in D, j \in \mathbf{N} \text{ s.t. } p_i^{i+k-1} = d_j^{j+k-1} \right\}.$$

**Probe nucleation complex algorithm** (Statement of the problem)

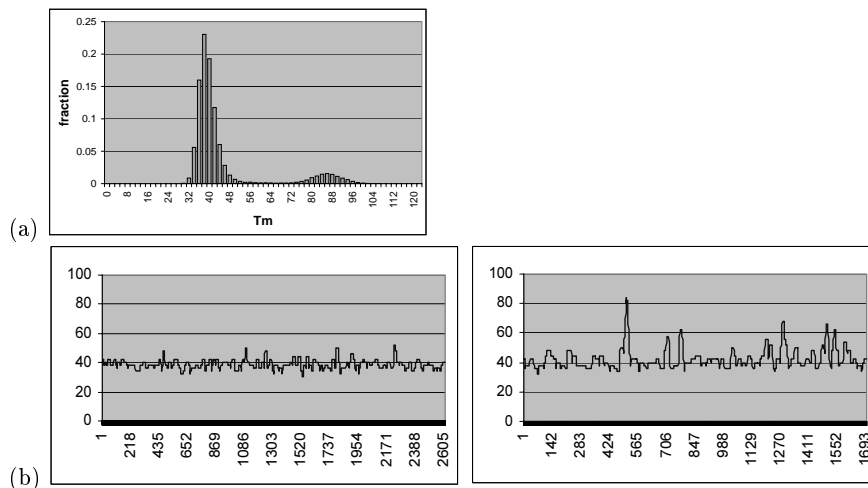
**Input:** The target gene  $g$  (of length  $m$ ), the background message  $D$  (of length  $N$ ), a probe length  $l$ , a melting temperature threshold  $t$ .

**Output:** A set of proximities of the candidate probes  $\{f(g_i^{i+l-1}, D) : 1 \leq i \leq m-l+1\}$ .

$T_m$  of short sequences may be approximated using different thermodynamic models (2-4 rule [15] nearest-neighbor [16] are examples). It should be noted that the expected size of the longest perfect match in a probe of length  $l$  is of order  $\log_4(lN)$  which, for  $l = 30$ ,  $N = 3 \cdot 10^6$  is  $\log_4(9 \cdot 10^7) > 13$ . As a consequence, an efficient indexed search with a large seed size may be used to accurately locate all possible locations of nucleation complex formation, followed by a rigorous calculation of  $T_m$  for these locations using the desired thermodynamic model.

<sup>7</sup> "... The process begins by the formation of a transient nucleation complex from the interaction of very few base pairs. Duplex formation proceeds, one base pair at a time, through a zippering process." [14]

Figure 4a depicts the distribution the proximities of candidate probes, in  $T_m$ , to the entire yeast transcriptome using the simple 2-4 model for  $T_m$  estimation. It is observed that this distribution is bimodal, composed of a distribution around  $T_m \approx 40$ , accounting for the expected probe proximities (13 bases with an average  $T_m$  of 3), and a second wider distribution around  $T_m \approx 90$ , for totally unspecific probes. From this distribution we may deduce that requiring a proximity threshold of  $T_m > 50$  should allow identification of unacceptably unspecific probes. Figure 4b illustrates probe proximity maps for two different genes.



**Fig. 4.** (a) Distribution of probe proximities (in  $T_m$ ) according to an arbitrary selection over a genomic database (200 ORFs of *S. cerevisiae*), for probe of length  $l = 30$ . (b) Distribution of probe proximities (in  $T_m$ ) for two different ORFs: GAL4 - a gene with no specificity issues and GAL2 - a gene that is highly homologous to genes from the HXT family.  $X$  axis denotes position of probe along the sequence. Compare to specificity maps of the same genes depicted in Figure 3.

## 4 Application to the Yeast Transcriptome

In Section 3 an algorithm for efficiently mapping the specificity of probes against a background with known content and distribution was presented. In practice, complete information of this kind is still rare although it is rapidly accumulating. For the yeast *S. cerevisiae* the entire theoretical transcriptome has been extensively mapped and studied. Publicly accessible databases contain the complete set of theoretical transcripts [17] as well as substantial experimental information about their distribution [18]. We used this existing information to create a full set of probe specificity maps for the entire *S. cerevisiae* transcriptome.

As mentioned earlier, specificity is not the sole parameter for probe selection and therefore, for each transcript, rather than selecting the "best" probe, we calculated a map that quantifies the specificity of each candidate probe. For each transcript, this map can then be used as the basis for selecting highly specific probes, in conjunction with sensitivity parameters.

Figure 3 shows the specificity maps of three different genes, for seed value  $\xi = 5$ . Two typical specificity patterns are encountered: either a distribution of probe specificities that is similar to the theoretical distribution (e.g. GAL4) or a distribution of much lower values that arises from similarity between genes (e.g. GAL2). Many transcripts combine domains of both types - some parts of the sequence show low specificity against the background, while others are unique (e.g. GAL1). For these latter transcripts the process of mapping specificity is of extreme importance.

Probe specificity mapping for *S. cerevisiae* was performed using the following criteria:

- The abundance weighted threshold values used determined according to the abundance of the target  $a_t$  and that of the background transcript  $a_b$ . Specifically, we used (for the indicated  $a_t/a_b$  pairs):  $h/h - 6$ ,  $h/m - 4$ ,  $h/l - 3$ ,  $m/h - 7$ ,  $m/m - 5$ ,  $m/l - 3$ ,  $l/h - 7$ ,  $l/m - 6$  and  $l/l - 4$ , where  $h$  stands for high abundance ( $> 10$  transcripts per cell on average),  $m$  - medium abundance ( $> 1$  transcript per cell) and  $l$  - low abundance (all other transcripts).
- Since they are intended for use in gene expression profiling, the map for each transcript was created only for its 500 long 3'-terminus (for transcripts shorter than 500 bases, the entire transcript was considered).

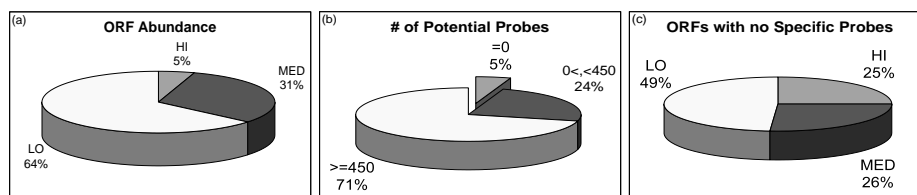
**Results** The complete set of specificity maps for the entire *S. cerevisiae* transcriptome will be available at our website. We present here a brief summary of the results of the mapping, and highlight some specificity issues.

Of the 6310 ORFs (theoretical transcripts) that appear in the *S. cerevisiae* genome databank [17] 71% (4475 ORFs) contain 450 or more valid probes, satisfying the aforementioned criteria. This bulk of the transcripts are those that have no significant specificity problems. The distance distribution for these transcripts is similar to the theoretical distribution described in Section 2. 24% of the ORFs (1497 ORFs) contain between 1 and 449 valid probes. These contain certain domains that are highly similar to other ORFs. Specificity mapping is important in this case, allowing the discrimination between specific and non-specific probes.

The remaining 5% of the ORFs (338 ORFs) are ones for which no satisfactory probe was found. Investigation of these ORFs revealed some explanations:

- A significant portion of the entirely non-specific ORFs are high-abundance transcripts (25% of the non-specific ORFs compared to 5% in the entire ORF population, see Figure 6). The non-specificity originated from gene duplicates that are either completely similar (e.g. TEF1/2 translation elongation factor, CUP1-1/2 copper-binding metallothionein) or almost completely similar (e.g. HHT1/2, HHT1/2 histone proteins, and many RPL and RPS ribosomal proteins). Although expression differences between such transcripts may

- be biologically significant (for example, if their regulation differs), the non-specificity of these transcripts is inherent to any hybridization-based assay.
- There are several groups of low-medium abundance transcripts that are very similar and therefore difficult to discriminate (e.g. HXT13/15/16/17 high-affinity hexose transporters, PAU1-7 seripauperin family). Functional difference between proteins in these groups is not always clear so the fact that they are indistinguishable may be problematic. Interrogating the expression levels of members of these sets requires further case-specific design.
  - High homology between an abundant and scarce transcript may cause the latter’s specificity to drop below the required criteria. A large percentage of non-specific ORFs appear in the group of unidentified ORFs, also referred to as hypothetical transcripts (7% non-specific, as opposed to 4% of the identified ORFs). Some of these may be dead genes, i.e. silent remnants of some partial duplication event. The hypothetical transcript YCL068C, for example, is very similar to a region within the abundant RAS guanyl-nucleotide exchange factor BUD5. In these cases non-specificity is may or may not an issue as the expression of the unidentified ORF may not be detected due to this similarity.



**Fig. 5.** Distribution of probe specificities for the 6310 ORFs of *S. cerevisiae*: (a) The overall distribution of ORF abundances, (b) the distribution of ORFs according to the number of their potential probes, (c) the distribution of probeless ORFs according to ORF abundance. Notice the large fraction of highly abundant ORFs that have no probes (25%) in relation to their fraction in the ORF population (5%).

## 5 Summary

Specificity issues that arise in the process of designing hybridization probes are discussed in this work. We start by providing rigorous approximation formulae for the distribution of the Hamming distance of an arbitrary word of length  $k$  (representing a candidate probe) to a uniformly drawn much longer sequence over the same alphabet. We then provide an algorithmic approach to efficient specificity mapping. The approach is then applied to the yeast transcriptome and results are compared to the stochastic model. A complete specificity map of the

yeast transcriptome was computed. For each ORF and each position we obtain an estimate of the potential cross-hybridization to the set of all other ORFs. When computing this potential we consider symbolic Hamming distance as well as relative abundance. Measurements for low copy messages are very sensitive to cross hybridization to high abundance messages. The majority of yeast ORFs follow the stochastic model. Message homologies result in specificity issues for many other ORFs. We expect such issues will need to be addressed case by case.

## References

1. A.P. Blanchard, and L. Hood, Sequence to array: Probing the Genome's Secrets, *Nature Biotechnology* 14:1649, 1996.
2. Y. Lysov, A. Chernyi, A. Balaev, F. Gnuchev, K. Beattie, and A. Mirzabekov, DNA Sequencing by Contiguous Stacking Hybridization on Modified Oligonucleotide Matrices, *Molecular Biology* 29(1):62-66, 1995.
3. E.S. Lander, Array of Hope, *Nature Genetics* 21:3-4, 1999.
4. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell* 9(12):3273-97, 1998.
5. M. Bittner et al, Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling, *Nature*, 406(6795):536-40, 2000.
6. T. R. Golub et al, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286(5439):531-7, 1999.
7. M. Mitsuhashi, A. Cooper, M. Ogura, T. Shinagawa, K. Yano and T. Hosokawa, Oligonucleotide Probe Design - a New Approach, *Nature* 367:759-761, 1994.
8. F. Li, G.D. Stormo, Selection of Optimal DNA Oligos for Gene Expression Arrays, *Bioinformatics* 17(11):1067-1076, 2001.
9. N. Hosaka N, K. Kurata, H. Nakamura, Comparison of Methods for Probe Design, *Genome Informatics* 12: 449-450, 2001.
10. A.D. Barbour, L. Holst, and S. Janson, Poisson Approximation, Clarendon Press, Oxford, 1992.
11. C. Stein, Approximate Computation of Expectations, Institute of Mathematical Statistics Monograph Series, Vol. 7, 1996.
12. M. Morris, G. Schachtel, and S. Karlin, Exact Formulas for Multitype Run Statistics in a Random Ordering, *SIAM J. Disc. Math.*, 6(1):70-86, 1993.
13. A.M. Mood, The Distribution Theory of Runs, *Ann. Math. Stat.* 11:367-392, 1940.
14. E. Southern, K. Mir, and M. Shchepinov, Molecular Interactions on Microarrays, *Nature Genetics*, 21(1):5-9, 1999.
15. T. Strachan and A.P. Read, *Human Molecular Genetics*, John Wiley & Sons, New York, 1997.
16. J. SantaLucia, A unified view of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-neighbor Thermodynamics, *PNAS USA* 95, 1460-1465, 1998.
17. Cherry, J. M., Ball, C., Dolinski, K., Dwight, S., Harris, M., Matese, J. C., Sherlock, G., Binkley, G., Jin, H., Weng, S., and Botstein, D., *Saccharomyces Genome Database*, <ftp://genome-ftp.stanford.edu/pub/yeast/SacchDB/>
18. F.C.P. Holstege, E.G. Jennings, J.J. Wyrick, T.I. Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young, Dissecting the Regulatory Circuitry of a Eukaryotic Genome, *Cell*, 95:717-728, 1998. <http://web.wi.mit.edu/young/expression/transcriptome.html>