Sensitivity and Capacity of Microarray Encodings

Max H. Garzon, Vinhthuy Phan,
Kiran C. Bobba, and Raghuver Kontham

Computer Science, The University of Memphis,
Memphis, TN 38152-3240, USA
{mgarzon, vphan, kbobba, rkontham}@memphis.edu

Abstract. Encoding and processing information in DNA-, RNA- and other biomolecule-based devices is an important topic in DNA-based computing with potentially important applications to fields such as bioinformatics, and, conceivably, microbiology and genetics. New methods to encode large data sets compactly on DNA chips has been recently proposed in (Garzon & Deaton, 2004) [18]. The method consists of shredding the data into short oligonucleotides and pouring it over a DNA chip with spots populated by copies of a basis set of noncrosshybridizing strands. In this paper, we provide an analysis of the sensitivity, robustness, and capacity of the encodings. First, we provide preliminary experimental evidence of the degree of variability of the representation and show that it can be made robust despite reaction conditions and the uncertainty of the hybridization chemistry *in vitro*. Based on these simulations, we provide an empirical estimate of the capacity of the representation to store information. Second, we present a new theoretical model to analyze and estimate the sensitivity and capacity of a given DNA chip for information discrimination. Finally, we briefly discuss some potential applications, such as genomic analysis, classification problems, and data mining of massive amounts of data in abiotic form *without* the onerous cost of massive synthesis of DNA strands.

Keywords: Data representation, Gibbs energy, h-distance, fault-tolerant computing, DNA chips, microarrays, genomic analysis, data mining, classification and discrimination.

1 Introduction

Biomolecular computing (BMC) was originally motivated by computational and engineering purposes. This endeavour would not be possible without some type of representation of data and information, directly or indirectly, onto biomolecules, both as input and as output in a computation. Virtually every application of DNA computing maps data to appropriate sequences to achieve intended reactions, reaction products, and yields. DNA molecules usually process information by intramolecular and (more often) intermolecular reactions, usually hybridization in DNA-based computing. The problem of data and information encoding

A. Carbone and N.A. Pierce (Eds.): DNA11, LNCS 3892, pp. 81–95, 2006.

on DNA bears an increasing interest for both biological and non-biological applications.

Most of prior work in this area has been restricted to the so-called *word design problem*, or even the encoding problem (Garzon et al., 1997) [10]. In this paper, however, we address a fairly distinct issue, herein called the *representation problem*. The problem is to find a systematic (i.e., application independent) procedure to map both symbolic (abiotic) and nonsymbolic (e.g., biological) information onto biomolecules for massively parallel processing in wet test tubes for real world problems. Mapping of non-biological information for processing *in vitro* is an enormous challenge. Even the easier direct readout problem, i.e., converting genomic data into electronic form for conventional analysis, is an expensive and time-consuming process in bioinformatics (Mount, 2001) [19]. Moreover, the results of these analyses are usually only available in manual form that cannot be directly applied to feedback on the carriers of genomic information.

Three properties are deemed critical for eventual success of a mapping algorithm/protocol. It must be (Blain and Garzon, 2004)[3]:

- **Universal**
  Any kind of symbolic data/pattern can be mapped, in principle, to DNA. Otherwise the mapping will restrict the kind of information mapped, and the processing capabilities in DNA form may be too peculiar or too constrained to be useful in arbitrary applications.
- **Scalable**
  Mapping can only be justified in massive quantities that cannot be processed by conventional means. Therefore it must be scalable to the tera-bytes and higher orders it will eventually encounter. Currently, no cost-effective techniques exist for transferring these volumes by manual addition and extraction of patterns one by one. Ordinary symbolwise transductions require manually manufacturing the corresponding DNA strands, an impossible task with current technology.
- **Automatic and high-speed**
  Manual mapping (e.g., by synthesis of individual strands) is also very costly timewise. An effective strategy must be automatable (from and back to the user) and eventually orders of magnitude faster than processing of the data *in silico*.

The purpose of this paper is to provide an analysis of a new approach recently proposed to represent data (Garzon & Deaton, 2004)[18, 8] that is readily implementable in practice on the well developed technology of DNA chips (Steckel, 2003) [21]. The method has the potential to represent in appropriately chosen DNA oligonucleotides massive amounts of arbitrary data in the order of tera- and peta-byte scales for efficient and biotechnologically feasible processing. Direct encoding into DNA strands (Garzon et al., 2003d) [18], (Baum, 1995) [1] is not a very efficient method for storage or processing of such massive amounts of data not already given in DNA form because of the enormous implicit cost of DNA synthesis to produce the encoding sequences, even if their composition were available. The more indirect, but more efficient, approach is reviewed in Section 2,

assuming the existence of a large basis of noncrosshybridizing DNA molecules, as provided by good codeword sets recently obtained through several sources (Deaton et al., 2002a; Garzon et Al, 2003) [7, 2]. The method appears at first sight to be plagued by the uncertainty and fuzzyness inherent in the reactions among biomolecular ensembles. In Section 2.2, we establish that these concerns are not justified by establishing, somewhat surprisingly, that it is possible to factor out noise and map symbolic data in a very "linear" fashion with respect to the properties of concatenation and set multiplicity on the symbolic side, and hybridization and amplification on the biochemical side. We further provide a preliminary experimental assessment of the sensitivity of the representation for problems such as recognition, discrimination, and classification. In Section 3, we also provide a theoretical analysis of the sensitivity and potential capacity of the method. Finally, in Section 4 we briefly discuss some advantages and potential applications, such as genomic analysis, classification problems, and data mining of massive amounts of data in abiotic form, as well as some problematic issues that require further study for wide implementation and application of the method.

2 Encoding Data and Information in DNA Spaces

The obvious method to encode data on DNA, namely a one-one mapping of alphabet symbols (e.g., bits) or words (e.g., bytes or English words in a dictionary) to DNA fragments could possibly be used to encode symbolic data (strings) in DNA single strands. Longer texts can be mapped homomorphically by ligation of these segments to represent larger concatenations of symbolic text. A fundamental problem with this approach is that abiotic data would appear to require massive synthesis of DNA strands of the order of the amount of data to be encoded. Current lab methods may produce massive amounts of DNA copies of the same species, but not of too many diverse species selected and assembled in very specific structures such as English sentences in a corpus of data (e.g., a textbook), or records in a large data warehouse. Even if the requisite number of species were available, the mapping between the data and the DNA strands is hard to establish and maintain, as the species get transformed by the reactions they must get involved in and they must be translated back to humanly usable expression.

An alternative more effective representation using recently available large sets of noncrohybridizying oligonucleotides obtainable *in vitro* (Chen et. al., 2005; Bi et. al, 2003) [4, 2] has been suggested in (Garzon and Deaton, 2004) [18]. We repeat next the basic definitions to make this paper self-contained. This method can be regarded as a new implementation of the idea in (Head et al., 1999; 2001) [16, 15] of aqueous computing for writing on DNA molecules, although through a simpler set of operations (only hybridization.) Since binary strings can be easily mapped to a four letter alphabet, we will simply assume that the data are given in DNA form over $\{a, c, g, t\}$. Representations using sets with crosshybridization

present are usually ambiguous and cannot be reliably used. More details on this point can be can be found in (Garzon and Deaton, 2004) [17, 18].

2.1 Representation Using a Non-crosshybridizing Basis

Let B be a set of DNA molecules (the encoding basis, or "stations" in Head's terminology (Head et al., 1999) [15], here not necessarily bi-stable), which is assumed to be finite and noncrosshybridizying according to some model of hybridization, denoted $|*,*|$ (for example, the Gibbs energy, or the h-distance in (Garzon et al, 1997) [10, 9]). We will also assume that we are provided some parameter coding for the stringency of reaction conditions τ (for example, a threshold on the Gibbs energy or the h-distance) under which hybridization will take place. For simplicity, it is further assumed that the length of the strands in B is a fixed integer n, and that B contains no hairpins. For example, if the h-distance is the hybridization criterion and $\tau = 0$, two strands x, y can only hybridize if they are perfectly complementary (i.e., $h(x, y) \leq 0$), so a maximal such set B can be obtained by selecting one strand from every (non-palindromic) pair of Watson-Crick complementary strands; but if, on the othr hand, $\tau = n$, the mildest hybridization condition, any two strands can hybridize, so a maximal set B consists of only one strand of length n, to which every other strand may hybridize without further restrictions. Let $m = |B|$ be the cardinality of B. The basis strands will also be referred as *probes*. For easy visualization, we will assume in the illustrating examples below that m is a perfect square $m = 36$ and that the base set of probess has been affixed onto a DNA chip.

Given a string x (ordinarily much longer than the probe length n and even perhaps the number of probes m), x is said to be *h-dependent on* B is there is some concatenation c of elements of B that will hybridize to x under stringency τ, i.e., such that $|x, c| \leq \tau$. Shredding x to the corresponding fragments according to the components of c in B leads to the following slightly weaker but more manageable definition. The *signature* of x with respect to B is a vector X of dimension m that is obtained as follows. Shredding x to $|x|/n$ fragments of size n or less, X_i is the number f of fragments of x that are within threshold τ from a strand i in B, i.e., such that $|f, i| < \tau$. The value X_i will thus be referred to as a *pixel* at probe spot i. The input strands x will also be referred as *targets*.

The only difference between a DNA-memory device and a DNA microarray is that the spots on the microarray consist of carefully chosen non-crosshybridizing DNA *basis* oligonucleotides rather than entire genes. Signatures can, however, be just as easily easily implemented in practice using currently available microarray technology.

For practical applications, a number of questions arise about this representation. First, the vector X may appear not to be well-defined, since it is clear that its expression depends on the various ways to find matching segments c in the input target x, the basis strands, and their concentrations. To start with, the number r of strands per spot, here called the *resolution*, can be varied at will and so change the intensity of each pixel and the resolution ability of the representation to distiguish various inputs. To avoid some of these techincal difficulties,

we will assume a relatively low resolution ($r = 6$ in the experiments below and $r = 1$ in the theoretical analysis of capacity.) On DNA chips, this resolution can be as high as the concentration (number of strands) of the basis strands (in solution), or as large as the number of strands per spot (on a chip.) More seriously, however, is the inherent uncertainty in hybridization reactions that make a signature dependent on the specific reaction conditions used in an experiment to "compute" it. From previous results in (Garzon and Deaton, 2004) [18], it is known that this problem disappears if a noncrosshybridizing set of high quality is used for the basis set. Experimentally, the signal to noise-ratio (precisely defined below) in the signature (given by the pixelwise ratio of signature signal to standard deviation of the same variable over all runs of the experiment) appears to be maximum. The hybridization likelihood between any pair of strands in a noncrosshybridizing set is minimized or even eliminated (by setting an appropriate stringency condition τ), regardless of the strands involved, the essential reason being that *a given fragment will can then only hybridize to at most one probe*. By assuming that either the test tube is small or that the reaction time is long enough that all possible hybridizations are exhausted within the experiment's time regardless of "kinetic bottlenecks", the basic problem thus becomes that of determining the set of possible signatures one may obtain by shredding the input in different ways, or even by using on different basis set.

In order to shed light on these questions, we performed a series of experiments with six target plasmids (described below) and three basis sets of different noncrosshybridizing qualities. The first set, $H40$, was obtained by randomly generating 40−mers and filtering out strands that whose h-distance is less than a given threshold ($\tau = 19\%$ of the shorter strands.) The second set, Ark, was obtained bottom up, by concatenating pairs of 20−mers randomly chosen from a set of 20−mers obtained by similar filtering and adding the resultant strand to the current membes of the set if its h-distance is greater than or equal to τ. The original set of 20−mers was obtained by using a more sophisticated genetic algorithm search using a Gibbs energy model (Deaton et al., 2002) [6] as fitness function. The third set, Hyb, was obtained by concatenating 40−mers from $H40$ and 20−mers from Ark and again adding the resultant strand to the set if its h-distance is greater than or equal to threshold h-distance ($\tau = 29$.) The noncrosshybridizing quality of these sets is high, as measured by the pairwise Gibbs energy of strands in the sets shown in Fig. 1.

Once the basis set and the reactions conditions have been optimized, the most important question remains, i.e., how *unique* is the signature for a given target x? To gain some insights into this question, six(6) large plasmids of lengths varying between $2.9K$ and $3.2K$ bps were chosen for targets and shredded into fragments of size 35 bps or less. Regarding the protocols as a stochastic process, experiments were conducted *in simulation* to obtain their signatures on a basis B as described above. Each experiment was run 10 times in a tested simulation environment, *Edna* (Garzon and Blain, 2004) [17] and (Garzon and Rose, 2004) [13]. As expected, we obtained a range of different signatures on different runs. Therefore, to make this concept precise, it is necessary to re-define a signature

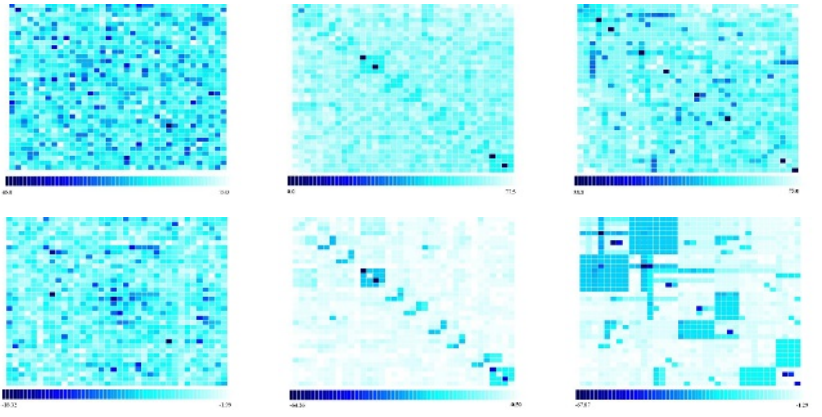

Fig. 1. Noncrosshybridization quality of a selection of three basis sets H40 (left column), Ark (middle column), and Hyb (right column) measured by the combinatorial h-distance (Garzon et al, 1997) [10] (the top row), and, the Gibbs energy model of (Deaton et al., 2002) [6] (bottom row). Their quality is high since lighter colors represent pairs far apart in hybridization distance or Gibbs energy (which is shown normalized to a comparable scale), i.e. lower hybridization affinity.

as a *sphere* in a high-dimensional euclidean space of dimension m (the number of spots on the microarray, i..e, number of noncrosshybridizng strands in the basis set.) The center of this sphere (below called the *ideal point signature*) is the componentwise average in mD-euclidean space of the outcomes of all possible point signatures obtained in running an experiment to find the signature. The radius of the sphere will be some measure of the variability of the all possible point signatures obtained in a given set of conditions. Here we use the *average euclidean distance* (i.e., the L_2-average) of all possible point signatures to the ideal signature.

With this definition of a signature as a sphere in mD-Euclidean space of radius given by the average distance from the ideal point signature, the problem of translating arbitrary data is resolved. We will refer to this sphere as the *volume* signature to distinguish it from the point signatures in the original definition. Examples can be seen in Fig. 5 (left). Fixing a basis set B, every target x determines a unique (volume) signature.

2.2 Sensitivity and Robustness of the DNA-Chip Representation

The critical question now about the signature of a given target x is the amount of information it contains, particularly to what extent it determines the target x uniquely, or, at least, whether it can distinguish it from other input targets. In this section we address these questions.

How much information about the target x does its volume signature provide? A comparison can be made using the so-called chipwise SNR (Signal-to-Noise Ratio) defined as follows. For each pixel X_i, SNR_i defined as the ratio of the

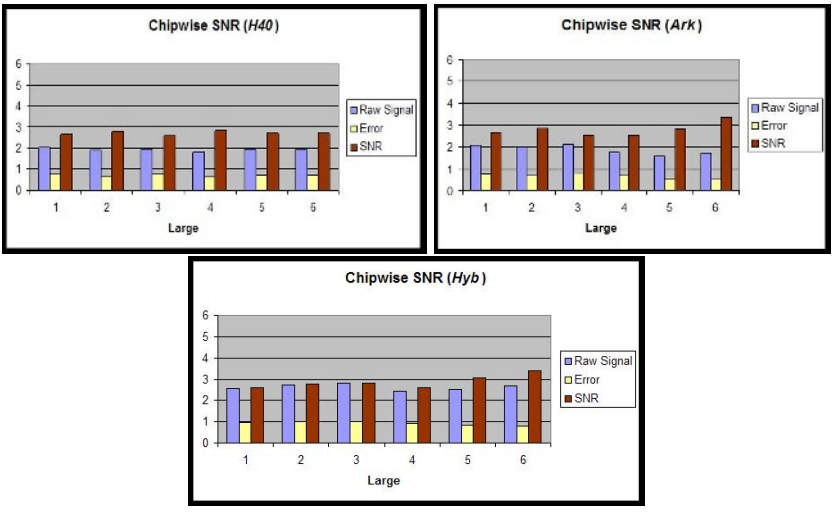

Fig. 2. Signal to Noise ratios (SNR) in six experiments with plasmids genomes over three sets H40 (top left), Ark (top right), Hyb (bottom) of various noncrosshybridization qualities

pixel's average value divided by the standard deviation of the same random variable X_i. The SNR of the target x (with respect to a given basis) is given by ratio of the L_2-average of the pixelwise signals divided by the L_2-average of pixelwise standard deviations. Fig. 2 shows the chipwise SNR comparison for all plasmids used in the experiments. The SNRs for the chosen plasmids are shown in Fig. 2. Some of them can be clearly distinguishable even if we just look at their SNRs alone, although it is too raw an average to expect full distinction among all plasmids. Nonetheless, the SNR gives a sense of the sensitivity of this representation.

There are other factors determining the radius of a volume signature that impact the variability of the representation. It is clear that slicing the input x into different fragments might change its volume radically, and that, conversely, reassembling the fragments in a different order may yields the same representation for a different input x'. How much does the representation depend on the lengths of the shredding x into pieces? The results described next provide an intuition on how Euclidean spheres radii change in representation signatures across a range of plasmid sizes ($2.9K$ to $3.2K$). Again, all experiments for sensitivity were performed ten times. Only results on the $H40$ probe set are shown below.

Fig. 3 shows the variations in the signature's radius obtained by varying the lengths of the the fragments shredding the target x. The radius increased for smaller fragments (15-25bp) compared to the original fragment size (25-35bp.) This increase is to be expected because more fragments are availability for hybridization, which results in higher signal and proportionately higher variability. The higher the standard deviation the bigger the radius of euclidean sphere.

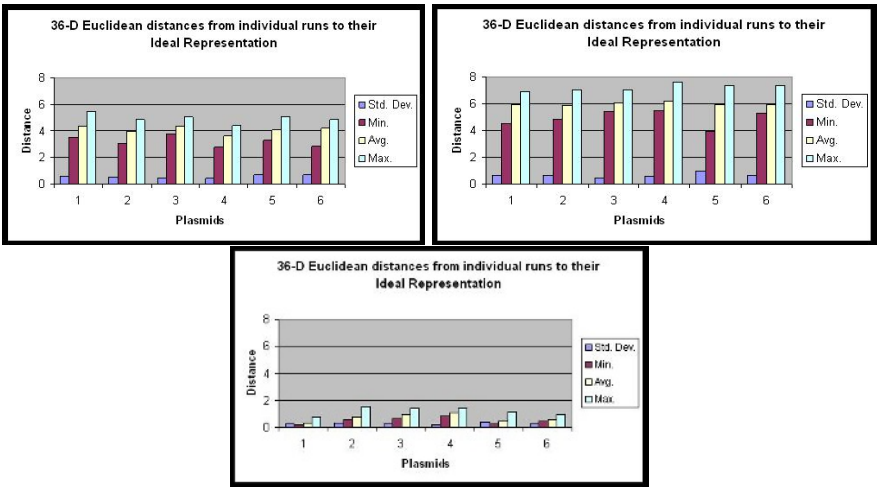

Fig. 3. Volume signature variability for original fragments of $25-35$bps (top left); small fragments of $15-25$bps (top right); and large fragments of $35-45$bp (bottom). Larger fragments yield a crisper signature (smaller radius) while shorter fragments yield fuzzier signatures (larger radius).

The converse argument can be given to explain the decrease in radius with large fragments (35-45bp).

Further experiments were performed to determine the sensitivity of the signature through contamination of targets in several ways. The contamination will be referred to as "noise." The noise introduced into original plasmids was of three types. The results described next provide a quantitative idea of the change expected in sensitivity of the signatures for plasmid 1. The target plasmid 1 was varied by introducing three types of noise:

- *Substitution*: Plasmids fragments are replaced by other random fragments of equal length;
- *Addition*: Random fragments were inserted in the plasmid;
- *Reduction*: Random fragments were removed from the plasmid.

Fig. 4 shows the variations in signatures of the resulting plasmid targets. The signatures' radii do not change much with substitution noise regardless of the amount substituted. However, the radii increased with increase in noise in the case of added noise and radii decreased with increased reduction noise. This behavior is similar for small and large fragments. This is additional evidence of sensitivity of the volume signature to changes in the length of and number of target fragments.

In order to determine the robustness of the representation, i.e., how much change must be made to a target for it to produce a different volume signature, we used the so-called *overlap* of volume signatures. This measure attempts to

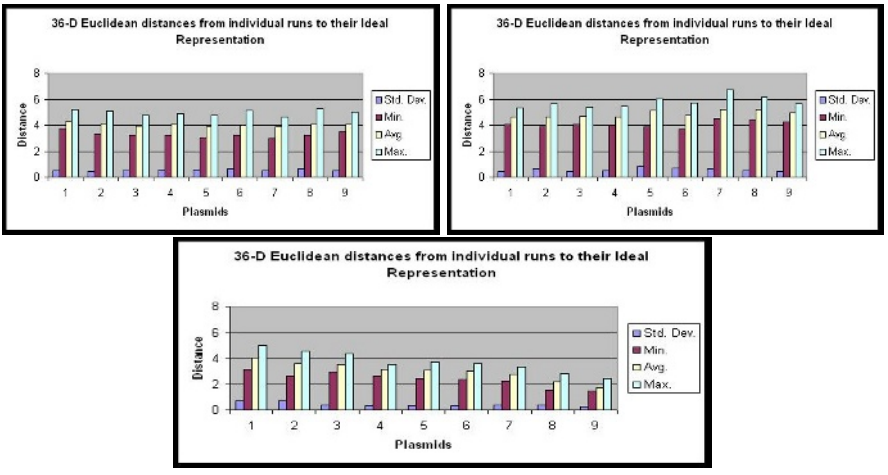

Fig. 4. Volume signature variability for target basis H40 for noise that has been substituted (top left), added (top right) or reduced (bottom). Volume signatures are also sensitive to changes in the length of and number of residues in the probe. However, the radii vary in proportion to noise. This behavior is similar for small and large fragments.

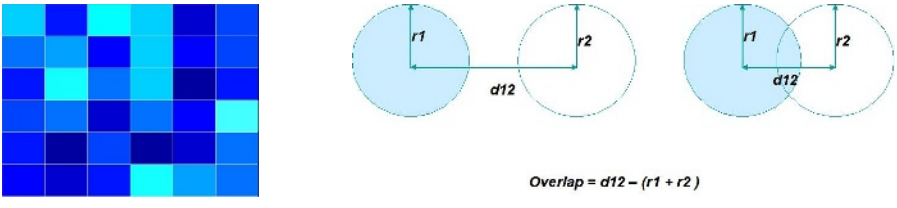

Fig. 5. The ideal representation of plasmid 1 (left). The overlap between two representations (spheres) is the excess (or defect) of the distance between ideal representations and the sum of the radii of the individual signatures. If the overlap is positive, the volume signatures do not intersect (middle), while they will if the overlap is negative (right).

capture the displacement in the ideal representation from its original parent with various types of noise, as shown in Fig. 5. *Overlap* is the difference between the distance between ideal representation and the sum of the average radii of their volume signatures. Fig. 6 shows the euclidean distances traveled from the ideal signature by variation of plasmid 1. Increasing substitution noise smoothly shifts the ideal signature but maintains overlap up to 60%. Only at 70% does the volume signature become nearly disjoint. With added noise, the threshold for the same phenomenon is about 90% noise, and with reduced noise it is about 60% noise. An overlap distance of -1 can be considered enough for two spheres to separate. So, it can be concluded that representations are sensitive to noise from Fig. 6 as the distance to original ones increase with increase in noise.

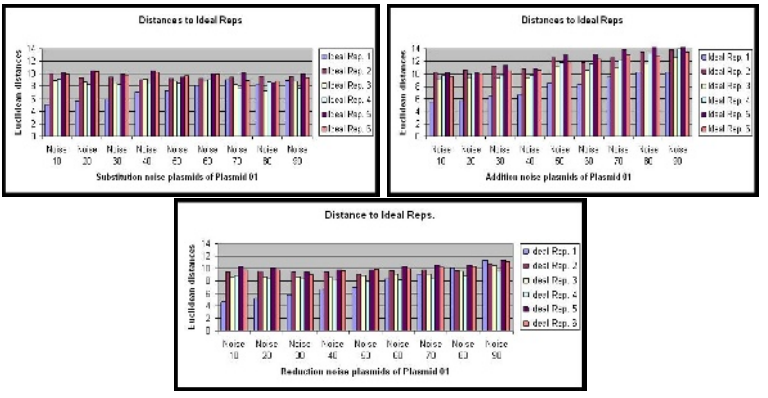

Fig. 6. Volume signatures are robust. It requires 70% for substitution noise (top left) and reduction noise (bottom) for a probe to become closer to others away from itself, while it remains closest to the original even with 90% added (top right) straneous fragments.

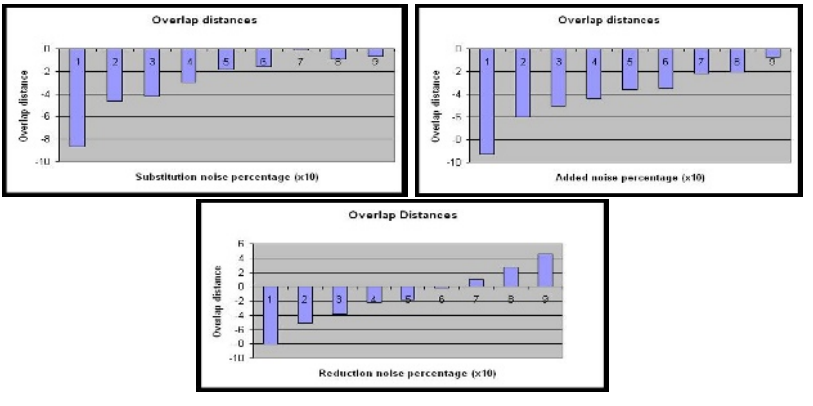

Fig. 7. Overlaps of volume signatures of noisy variations of plasmid 1 to its original volume signature. Substitution noise of 70% (top left) is required for the volume signature to become nearly disjoint. With added noise (top right), the threshold for the same phenomenon is about 90% noise. With reduced noise (bottom), the threshold it is about 60% noise. An overlap distance of -1 can be considered enough for two spheres to separate. Thus, representations are fairly insensitive to a small amounts of noise, while remaining sensitive to larger changes.

Fig. 7 shows a further analysis of the same experiment by considering the distances of the noisy plasmid 1 to the ideal signature of all six plasmids. The most interesting threshold is the amount of noise required for varying plasmid 1 to become closer to another plasmid than to its original. That number is 70% for substitution noise and reduction noise, but the noisy plasmids remain closest to

the original even under 90% added straneous fragments. This is is a remarkable robustness.

3 Theoretical Analysis of Sensitivity and Capacity of DNA-Based Chips

We now provide an abstraction of the concept of a signature in order to provide a theoretical model to estimate the capacity of DNA chips under optimal conditions. First, due to the fact that, under realistic conditions, it is infeasible to expose very long uncut copies of an input sequence to the chip, we assumed in the definition of signature that the targets are shredded by restriction enzymes into manageable fragments before they are exposed to the chip. To simplify the analysis in the theoretical model, however, we will assume that no shredding of targets will be carried out.

To justify this assumption, we observe that we can disregard all basis strands that are Watson-Crick complementary to the cleaving restriction sites used for shredding since hybridizations of targets to basis strands in the vicinity of the restriction sites will not be happen. Therefore, we can eliminate shredding if we guarantee that the basis set contains no restriction site used by shredding enzymes and still get an identical signature for the same target. Second, we will assume that basis strands float freely in solution instead of being affixed to a chip, which is justified given the nonhybridization property of the basis set. Third, we will also assume that a fixed concentration of basis strand and targets is placed in the tube. Thus, target strands are exposed, in principle, to hybridization of all basis strands at many places, and, consequently, many copies of the same basis strand may hybridize to several parts of the input sequence. Thus, even though target sequences can be arbitrarily long, there can only be a bounded number of point signatures, and so different targets may yield the same point signature. Under these assumptions, the volume signature produced by an uncut input target is essentially the same as the one produced by the the original definition above.

In this model, the chip capacity (i.e., the number of distinguishable target signatures) becomes a function of σ, *i.e.*, the total number of copies of all basis strands that hybridize to a target.. Realistically, when σ varies slightly, so does its capacity. Given a set $B = \{i_1, i_2, \cdots, i_k\}$ of *basis strands* and a target sequence X, its *signature* is $x_B = (x_1, x_2, \cdots, x_m)$, where x_i is the number of times basis oligo i hybridizes to (different parts) of X. Under these conditions, input targets X and Y are indistinguishable if and only if $x_B = y_B$, i.e. $x_i = y_i$, for all $1 \leq i \leq m$.

The basis B used to create a DNA chip relates to the capacity of the chip in interesting ways. We observed that the arguments in (Phan & Garzon, 2004) [20] show that the memory capacity of the noncrosshybridizing basis B is large if (a) its oligo distribution as substrings of the input sequences is as far from uniform as possible; and, (b) they *cover* the input targets as much as possible. Specifically, we found that

Proposition 1. *The probability of two different input sequences being indistinguishable from each other is*

$$P(X_B = Y_B | X \neq Y) = \frac{\binom{\sigma}{x_1, x_2, \cdots, x_k}}{k^\sigma} \leq \frac{2^{\sigma H(P)}}{k^\sigma} = \frac{1}{2^{\sigma(\log_2 k - H(P))}} \qquad (1)$$

where $\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} y_i = \sigma$, *and* $H(P) = -\frac{x_i}{\sigma} \sum_{i=1}^{k} \frac{x_i}{\sigma}$, *the Shannon entropy of the distribution of* B *in* X *(and* Y*).*

In other words, the capacity of the chip based on B is small if one of two conditions are true:
(1) σ is small, or (2) the distribution of the bases as substrings of the inputs sequences approaches random (i.e. $H(P)$ approaches $\log_2 k$). When B covers the input sequences completely, every substring of an input of the same length as the $|s_i|'s$ hybridizes to one of the bases, and consequently $\sigma \approx \frac{|X|}{|i|}$, where $|i|$ is the length of basis oligo i. Conversely, when B covers the input sequences sparsely, $\sigma \ll \frac{|X|}{|i|}$ and the probability of two different input sequences being indistinguishable increases.

Using these arguments, we can also provide a theoretical estimate of the capacity of the DNA chip for volume signatures as defined above. The limit of a DNA chip's capacity is the number of distinguishable signatures that the chip can possibly produce. Since the total number of occurrences of each basis strand (x_i's) in X adds up to σ, we have the following conditions:

$$\forall i, (x_i \geq 0), \qquad \text{and} \qquad x_1 + x_2 + \cdots + x_k = \sigma \qquad (2)$$

Using an elementary combinatorial argument, we can show that

Proposition 2. *The optimal capacity a DNA chip is* $\binom{\sigma+m-1}{m-1}$*, if defined as the maximum number of distinguishable point signatures.*

As mentioned above, it is not the case that exposing an input target a number of times will get an identical signature each time. In the current mode, where the chip is not affixed but in solution, this sensitivity to distinguish input is decreased. because the signatures of different but similar input sequences are likely indistinguishable. The sensivity of the chip can be collectively captured by two parameters r and r_σ, regardless of the sources of noise. The capacity of the chip is estimated indirectly via the size of a maximal set, called C, in signature space. This set C can be thought of as a *maximal* collection of centers of non-intersecting spheres with a fixed radius r. Hence, the sphere of radius r specifically captures the uncertainty of telling signatures of similar sequences apart; sequences whose signatures are within a radius r are not distinguishable. The other parameter, r_σ captures the fact that due to noise or other factors, even when the bases *cover* well input sequences, the number of basis strands hybridized to these inputs may not always be exactly σ. Hence, we assume that the total number of basis strands hybridized to the input sequences vary from $\sigma - r_\sigma$ to $\sigma + r_\sigma$. On these considerations, we have established the following estimate, where the set V consists of the signatures of all input targets that the chip could distinguish under the sensitivity parameter r.

Theorem 1. *The maximal set C of signatures that are distinguishable on a DNA-based (m, σ, r)-chip is of size $|C|$ bounded by*

$$\frac{|V|}{v(2r+1)} \leq |C| \leq \frac{|V|}{v(r)} \tag{3}$$

A full proof is omitted. Briefly, these bounds are obtained by determining the upper and lower bounds of a maximal code, in a similar fashion as the Hamming and Gilbert-Varshamov bounds, respectively. Intuitively, the input sequences in V include those whose signatures fall inside the hyperplane in equation 2 and those input sequences whose signatures fall within a distance r of the hyperplane.

Lemma 1

$$|V| = \binom{\sigma + m - 1}{m - 1} + \sum_{i=1}^{r_\sigma} 2 \binom{\sigma + i + m - 1}{m - 1}$$

$|V|$ is, however, not the same as $|C|$; i.e. it is not the capacity of the chip because two signatures within a distance of r from each other are not disintiguishable. To estimate $|C|$, we need to know, $v(r)$, the number of signatures inside a sphere of radius r.

Lemma 2

$$v(r) = 1 + \sum_{e=1}^{r} \sum_{i=1}^{min\{e,k\}} 2^i \binom{k}{i} \binom{e - i - 1}{i - 1}$$

Proof. A full proof is omitted for space reasons. Briefly, the sum accounts for all points at distance exactly e from a center, for $0 \leq e \leq r$. □

4 Conclusions and Future Work

This paper gives experimental (in simulation) and theoretical analyzes of a recenly proposed method (Garzon and Deaton, 2004) [18] to represent abiotic information onto DNA molecules in order to make processing data at massive scales *efficient* and *scalable*. The mapping is readily implementable with current microarray technology (Stekel, 2003) [21], bypasses synthesis of all but a few strands, and it's promising for the tera- and peta-byte scopes volumes required for a meaningful applications (more below.) Furthermore, we quantify the sensitivity of the representations and show that it can be made robust despite the uncertainty of the hybridization chemistry. Third, we show a theoretical analysis of the capacity of this type of representation to code information, as well as an information-theoretic estimate of the number of distinguishable targets that can be representated on a given chip under reaction conditions characterized by hybridization stringency parameters.

A direct application of this method in bioinformatics is a new approach to genomic analysis that increases the signal-to-noise ratio in microarrays commonly

used in bioinformatics. The method yields higher resolution and accuracy in the analysis of genomica data, and only requires some processing in what can be termed an "orthogonalization" procedure to the given set of targets/genes before placing them on the microarrays. These advantages may be critical for problems such as classification problems (disease/healthy data). More details can be found in (Garzon et al., 2005) [12].

Further applications can be expected in the analysis and data mining of abiotic data, whose representation is automatically defined with respect to a given basis set B. Given a noncrosshybridizing basis and adequate thresholds on the stringency of reaction condition and acceptable levels of variability of the representation (i.e., the capacity to distinguish inputs through their representations), the signatures of arbitrary inputs are completely determined and require no precomputation or synthesis of any DNA strands, other than the basis strands. In other words, this method provides a *universal* and *scalable* method to represent data of any type. For example, because of the superposition (linearity) property (module the variability implicit in the representation), a corpus of English text can be automatically encoded just by finding representations for the words in the basic vocabulary (words) in the corpus. Thereafter, the representation of a previously unknown piece of text can be inferred by superposition of the component words. There is evidence that these representations can be used for semantic processing of text corpora in lieu of the original text [11]. Given the newly available large basis sets [4, 5, 6] in the order of megasets, device with the ability to process data for information extraction appear now within reach in a relatively short time.

Acknowledgements

Much of the work presented here has been done in collaboration with a molecular computing consortium that includes Russell Deaton, Jin Wu (U. of Arkansas), Junghuei Chen, and David Wood (U. Delaware). Support from the National Science Foundation grant QuBiC/EIA-0130385 is gratefully acknowledged.

References

1. E. Baum. Building an associative memory vastly larger than the brain. *Science*, 268:583–585, 1995.
2. H. Bi, J. Chen, R. Deaton, M. Garzon, H. Rubin, and D. Wood. A pcr-based protocol for in vitro selection of non-crosshybridizing oligonucleotides. *J. of Natural Computing*, 2003.
3. Derrel Blain and M. Garzon. Simulation tools for biomolecular computing. *In: (Garzon and Rose, 2004)*, 3:4:117–129, 2004.
4. J. Chen, R. Deaton, M. Garzon, J.W. Kim, D.H. Wood, H. Bi, D. Carpenter, J.S. Le, and Y.Z. Wang. Sequence complexity of large libraries of dna oligonucleotides. *In these proceedings*, 2005.
5. J. Chen, R. Deaton, Max Garzon, D.H. Wood, H. Bi, D. Carpenter, and Y.Z. Wang. Characterization of non-crosshybridizing dna oligonucleotides manufactured in vitro. Proc. 8th Int Conf on DNA Computing DNA8.

6. R. Deaton, J. Chen, H. Bi, and J. Rose. A software tool for generating non-crosshybridizing libraries of dna oligonucleotides. pages 252–261, 2002. In: [14].
7. R.J. Deaton, J. Chen, H. Bi, M. Garzon, H.Rubin, and D.H. Wood. A pcr-based protocol for in vitro selection of non-crosshybridizing oligonucleotides. *In: (Hagiya & Ohuchi, 2002)*, pages 105–114, 2002a.
8. M. Garzon, K. Bobba, and B. Hyde. Digital information encoding on dna. *Springer-Verlag Lecture Notes in Computer Science 2590(2003)*, pages 151–166, 2003b.
9. M. Garzon, R. Deaton, P. Neathery, R.C. Murphy, D.R. Franceschetti, and E. Stevens Jr. On the encoding problem for dna computing. pages 230–237, 1997. Poster at The Third DIMACS Workshop on DNA-based Computing, U of Pennsylvania. Preliminary Proceedings.
10. M. Garzon, P.I. Neathery, R. Deaton, R.C. Murphy, D.R. Franceschetti, and S.E. Stevens Jr. A new metric for dna computing. *In: (Koza et al., 1997)*, pages 472–478, (1997a).
11. M. Garzon, A. Neel, and K. Bobba. Efficiency and reliability of semantic retrieval in dna-based memories. pages 157–169, 2003.
12. M. Garzon, V. Phan, K. Bobba, and R. Kontham. Sensitivity analysis of microaary data: A new approach. In *Proc. IBE Conference, Athens GA.*, 2005. Biotechnology Press.
13. M. Garzon and John Rose. Simulation tools for biomolecular computing. *Special Issue of the Journal of Natural Computing*, 4:3, 2004.
14. M. Hagiya and A. Ohuchi. In *Proc. 8th Int. Meeting on DNA-Based Computers.*, 2002. Springer-Verlag Lecture Notes in Computer Science LNCS 2568. Springer-Verlag.
15. T. Head, M. Yamamura, and S. Gal. Aqueous computing: Writing on molecules. 1999. Proceedings of the Congress on Evolutionary Computing (CEC'99).
16. T. Head, M. Yamamura, and S. Gal. Relativized code concepts and multi-tube dna dictionaries. In *Finite vs Infinite: COntrobutions to an eternal dilemma (Discrete math and Theoretical Computer SCience)*, pages 175–186, 2001.
17. Garzon M, D. Blain, and A. Neel. Virtual test tubes for biomolecular computing. *In: (Garzon and Rose, 2004)*, 3:4:460–477, 2004.
18. Garzon M and R. Deaton. Codeword design and information encoding in dna ensembles. *J. of Natural Computing*, 3:4:253–292, 2004.
19. D. Mount. Bioinformatics: sequence and genome analysis. *Spring Harbor Lab Press, MD*, 2001.
20. V. Phan and M. Garzon. Information encoding using dna. *Proc. 10th Int Conf on DNA Computing DNA10*, 2004.
21. D. Stekel. *Microarray Bioinformatics*. Cambridge University Press, 2003.