Retrieval Efficiency of DNA-Based Databases of Digital Signals

Sotirios A. Tsaftaris, *Member, IEEE* and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—Using DNA to store digital signals, or data in general, offers significant advantages when compared to other media. The DNA molecule, especially in its double stranded form, is very stable, compact, and inexpensive. In the past, we have shown that DNA can be used to store and retrieve digital signals encoded and stored in DNA. We have also shown that DNA hybridization can be used as a similarity criterion for retrieving digital signals encoded and stored in a DNA database. Retrieval is achieved through hybridization of 'query' and 'data' DNA molecules. In this paper, we present a mathematical framework to simulate single query and parallel query scenarios and to estimate hybridization efficiency. Our framework allows for exact numerical solutions as well as closed form approximations under certain conditions. Similarly to the digital domain, we define a DNA signal-to-noise ratio (SNR) measure to assess the performance of the DNA-based retrieval scheme in terms of database size and source statistics. With approximations, we show that the SNR of any finite size DNA-based database is upper bounded by the SNR of an infinitely large DNA-based database that has the same source distribution. Computer simulations are presented to validate our results.**

***Index Terms*—DNA, signal processing, DNA-based digital signal processing, simulations, hybridization, modeling.**

I. Introduction

ADLEMAN demonstrated the computational capacity of DNA, by solving a specific combinatorial problem, the Hamiltonian path problem, applying principles of DNA chemistry [1]. Baum [2] claimed that it is possible to build a DNA database that encodes digital instead of genetic information. He argued that this database could have enormous storage capacity and can retrieve information based on content, very similar to how the human brain works.

Using DNA to store digital signals or data in general offers significant advantages when compared to other media. The DNA molecule, in its double stranded form, is very stable, compact, and inexpensive. Double stranded DNA, when preserved appropriately can last many years [3]. (If suspended in aqueous environments, DNA is susceptible to hydrolysis, the process of reacting with water.) A database can be easily and economically replicated by Polymerase Chain Reaction (PCR). Searching the database can be implemented with a plethora of techniques. Given a query, DNA hybridization provides an efficient way to search for similar molecules in the database. In digital databases the search time typically increases with the size of the database. However, in DNA databases when hybridization is used as a search mechanism, the querying time does not depend on the size of the database. This is because DNA kinetics depend on relative concentrations and molecular diffusion but not on the number of different molecules. Furthermore, parallel queries can take place thus improving the throughput of searching (using microarrays for example, a technology that is now commonly used in genomic analysis).

Taking into consideration all of the aforementioned qualities, it comes as no surprise that DNA has been considered a great candidate for storage of digital data [4]–[8], and even biological data [9]. Motivated by this fact, in our work we use DNA to store digital signals and hybridization to search through the database [10]–[13]. Overall, DNA-based storage can be considered an organic based approaches to digital signal processing [14].

Let us first consider the problem of searching in a digital database of digital signals. Consider a set of M digital signals each of length k and each entry of which is an r-bit integer. Consider also a vector q_d that contains $k_Q < k$, r-bit integers. The problem at hand is to find out whether q_d can be found in the database. Traditionally a matching criterion must be defined that measures the similarity between the query and the digital signal [15]. Overall the retrieval goal is to provide (a) a yes/no answer of whether a match has been found and (b) the locations and the identities of the signals where matches have occurred.

First step in building a DNA equivalent of a digital database, is to encode digital signals into DNA sequences, which is also known as the codeword design problem. The success of any DNA computation depends largely on the codewords. In our problem, the encoding has to be such that it enables content-based searches but limits retrieval errors. Furthermore, the encoding scheme needs to account for the presence of noise and allow for imperfect matches. To accomplish this, we introduced the Noise Tolerance Constraint [10].

The second step is to decide on the structure of the database elements. Each molecule is considered to be a database element; the database consists of a collection of multiple copies of database elements, which store the signal information. Usually each element has a unique index block (or address), which uniquely identifies it and/or enables the retrieval of a usually larger information (data) carrying block. There are many different ways to design database elements, each of which has unique properties and characteristics (a comparison is presented in [11]). To construct a DNA database blocks of digital information are converted into DNA sequences using the digital to DNA encoding and the molecules are synthesized.

In place of experimental verification of the proposed DNA database scheme we have developed a mathematical model that can simulate hybridization reactions between query and target molecules [16]. We implemented this framework and tested it on a small scale database to show that hybridization efficiency is inversely proportional to the mean squared error (MSE) of the encoded signal values [12], [13].

The main contributions of this article are: (a) the performance study of very large size databases and (b) the extension of our framework in modeling parallel query retrieval. Similarly to the digital domain, we define a signal to noise ratio (SNR) metric to quantify the performance of the DNA retrieval scheme in single and parallel querying situations. Our framework allows for exact numerical solutions as well as approximations under certain conditions. With approximations, we show that the SNR of a DNA database is upper bounded by the SNR of an infinitely large one with the same source distribution. This shows that in terms of retrieval accuracy, there is actual performance gain as the size of the database increases. Furthermore, we show that the same bound holds for parallel query retrieval when a certain laboratory protocol is followed.

Manuscript received July 3, 2008; Revised April 4, 2009.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: stsaft,aggk@eecs.northwestern.edu.)

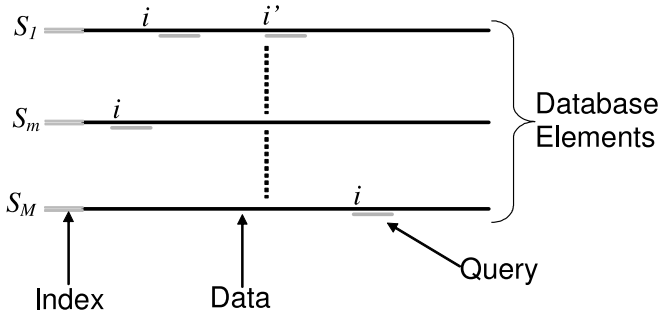

Fig. 1. Illustration of hybridizations between query and database elements. The variable i indicates location.

This paper is organized as follows. In Section II we sketch the characteristics of the equivalent DNA database system that can store digital signals. The framework for modeling of single and parallel query searches, under equilibrium assumptions, in DNA databases and performance evaluation using an SNR metric are presented in Sections III, and IV, respectively. Our study on the SNR of an infinitely large database is presented in Section V. Simulation results are given in Section VI. Finally in Section VII conclusions are given along with possible future extensions and applications.

II. System Description

A number of constraints have been proposed in order to satisfy a specific input problem but also limit errors that arise either naturally, or technically [17]–[20]. In our case the problem translates into finding N DNA sequences or words $w_i, i = 0, \ldots, N-1, (= 2^r-1)$, from the quaternary alphabet A, T, G, and C, each l bases long, capable of encoding integer signal values $i = 0, \ldots, N-1$. In short, this is a look up table that matches signal values, i, to the fixed length sequences w_i.

For the problem of encoding signals careful consideration is necessary to account for the noise in the signals. Digital signals are encoded satisfying the Noise Tolerance Constraint [10], that is codewords corresponding to integers numerically close to each other are similar (they have similar thermodynamic characteristics), while integer values far apart have codewords rather dissimilar. More details on the codeword design and a set of words capable of encoding 5-bit signals are given in the Appendix. We should note that most aspects of the following analysis hold for all hybridization based retrieval schemes and are independent of the codeword design methodology.

We assume that DNA sequences inside a database are constructed as shown in Fig. 1, following Baum's model [2]. Readers are referred to [11] for a discussion and comparison of other possible designs. For each database element $S_\jmath$, with concentration $C_\jmath$, the double-lined gray part is the index that identifies the data, which are shown as solid black lines. Data are concatenations of DNA words w_i, $i = 0, \ldots, N-1$. The index part of different elements should be very dissimilar. Assume that we have M digital signals and hence M database elements of length $(L + IN)$ bases, where IN, L is the index and data length, respectively.

The system is described with the following parameters and inputs:

1) M database elements $S_\jmath$, each of concentration $C_\jmath$ and sequence information $s_\jmath$ each of length L, $\jmath = 1, \ldots, M$.
2) A query Q, shown in Fig. 1 as a solid gray line, of concentration $|Q|_o$ and sequence information s_Q of length l_q.
3) Reaction parameters: temperature T and salt concentration $\left|Na^{++}\right|$.

To retrieve information from the database query molecules are synthesized. Queries are signal segments of interest. The query signal is encoded using the same look-up table, but the complementary sequence is synthesized and introduced in the solution. The query molecules will hybridize to complementary molecules in the database as seen in Fig. 1. There is a plethora of laboratory techniques to assist in the hybridization and filtering process (eg., affinity purification, FACS) and some even allow for parallel query searches. However the focus of this article is to quantify the percentage of correct retrievals that is the percentage of query molecules that hybridize to desired targets versus erroneous ones.

Hybridization between molecules is a random process and the probability of two molecules hybridizing is a function of concentrations, thermodynamic strength of their chemical bond, T and $\left|Na^{++}\right|$ [21]. Therefore, it is critical to quantify the percentage of desired hybridizations over the complete ensemble of hybridizations. Consequently, a SNR metric can be defined, where signal is considered to be the concentration of events corresponding to desired hybridizations and noise the concentration of undesired ones.

III. Modeling Hybridization Reactions

In this section we present a framework for simulating query searches in DNA databases. We first explore single query searches and offer metrics of their performance, such as error estimates and scalability with respect to different query and database concentrations, source statistics, and number of database elements. We then extend our analysis by simulating multi-query environments where parallel queries take place. We use a 2nd order thermodynamic equilibrium model that provides a tractable computational solution, which can be further simplified via linearization to provide a closed form solution. Despite the model's simplicity, it provides a best case upper bound for the performance of the database.

To aid in our presentation we introduce the notion of fragments as in [22]. A fragment $F_{i,p}^\jmath$ represents the sequence information of a database element $\jmath$ at location i of length p with concentration $\left|F_{i,p}^\jmath\right|$. It is clear that $F_{i,p}^\jmath \subseteq s_\jmath$. Furthermore, it is apparent that in our case the initial concentration of each fragment is equal to the concentration of each database element, that is $\left|F_{i,p}^\jmath\right|_o = C_\jmath$.

Subsequently we denote the query fragment complexes as $QF_{i,p}^\jmath$. Such complexes are illustrated in Fig. 1 at various locations. The complexes can have rather elaborate geometric structures (also referred to as secondary structures) and a number of modeling approaches have been considered [21], [23]–[28]. We will also assume that only linear query-fragment complexes of length $p = l_q$ are formed, and we will thus drop p from our notation. This assumption implies that complexes will only have internal mismatches, no loops, and no dangling ends. Furthermore, we assume that the index and the query will not cross-hybridize, due to the way the index was designed.

Data are concatenations of codewords w_i, $i = 0, \ldots, N-1$ of length l, and queries are concatenations of complements of codewords. In contrast to the analysis presented in [16] we assume that the query is a single codeword $Q = w_i^C$ and that only perfectly aligned complexes are formed; therefore each fragment is a word $F_i^\jmath = w_j$. This will allow us to relate hybridization efficiency and performance with database size, a relationship that could not be analytically derived without the above assumptions.

Under these assumptions query fragment complexes $QF_i^\jmath$ are actually codeword pairs $w_i^C w_j$. Since we have M database elements of concentration $C_\jmath$ and each database element is L bases long, the total number of complexes N_T is equal to $N_T = M\frac{L}{l} = M \cdot k$. However, there exist only N^2 distinct complexes (N molecules and N Watson Crick complements), and if $N_T < N^2$ multiple occurrences of the complexes occur. Hence for a given w_i^C, we have

N hybridization reactions of the form,

$$w_i^C + w_j \underset{K_r}{\overset{K_f}{\rightleftharpoons}} w_i^C w_j, \quad \forall j \in [0, N-1]. \tag{1}$$

The parameters K_f and K_r are called respectively the forward and reverse rate constants, and they depend on environmental parameters and laboratory settings. They are usually difficult to estimate since they require a plethora of laboratory experiments and they are not universal. Therefore, usually an equilibrium analysis that does not model the dynamic behavior is sought after.

A. Model Validity

The assumptions made in the section above that resulted in (1) are realistic, since our codewords were designed to limit the formation of secondary structures and the index words are designed to be highly dissimilar with the codewords [10], [29]. However, it is true that due to the stochastic nature of hybridization reactions such errors will happen with small probability. Accounting for those events and for all possible secondary structures is an exhausting exercise and requires numerous computing cycles. For example the formulation in [30] provides an elegant and computationally efficient algorithm with $O(N^4)$ complexity, where N is the number of complexes considered. However, the algorithm assumes equimolar concentrations of all the strands involved.

In our analysis we want to include the initial concentration of the species as a system parameter. We risk however, the validity of our thermodynamic models since the 2nd order equilibrium reaction model assumed is not entirely accurate when one of the species is in excess (such as when the database reaches infinity) [23], [31]. In our case however, the species that are in excess are the database elements and not the query. We are interested in the complexes where the query is involved and not in the unimolecular secondary structures that the database elements might form. We have examined concatenations of our codewords, and we have verified via the algorithm in [30] that they do not form competing secondary structures against the desired query and database element complexes when the query is diluted. (The detailed presentation of the verification goes beyond the scope of this article.) As we will show in more detail below, when the query is diluted the competition for the query molecules is large, and as long the query molecule does not form any secondary structures that inhibit query fragment complexes, our analysis is still valid.

To this effect, our numerical analysis provides an upper bound of performance via the solution of linear equations that allows the user to reach a fast conclusion about the quality of the design. In the Virtual Test Tubes of [32], [33] hybridization affinity is approximated with a modified Hamming distance metric, namely the H-measure. It was shown that even this simplified representation provides simulation results that are close to experimental reality. In [34] statistical thermodynamics are used to study the interaction of hybridizing nucleic acids and reach similar conclusions as the ones presented here.

A multiplex qPCR assay, used in measuring gene expression, can be considered as a parallel query retrieval in a DNA database (the target genomic material). Despite the careful design of primers (probes), most of the assumptions above are violated. There will be secondary structure in the targets or target to target hybridization that might inhibit probe-target hybridization. In this case, it is necessary to include all possible secondary structures and go beyond the two state model (eq. 1) and consider the competition of all these states and molecules in a multiplex-multi-state model [35]. A multiplex approach has been followed by Horne and colleagues [35] and it offered valuable conclusions to the effect of competition between cross-hybridized products and secondary structures on desired hybridization based both on kinetic and equilibrium analysis. They also compared with other methods, multiplex or multi-state, in the literature. We should highlight that when the authors made similar assumptions to ours, the reaction equations and equilibrium analysis are the same; however, we offer a different computational solution and we even provide an approximate closed form solution.

B. Estimating the concentration of query fragment complexes

The objective of this section is to estimate the concentration of complexes $w_i^C w_j$, denoted by $\left|w_i^C w_j\right|$, in equilibrium by assuming that all database elements have equal concentration, that is $C_j = C$. We will proceed to find $|w_i^C w_j|$ following similar steps as in [16]. Under an equilibrium assumption, the differential equations that describe the mass action equations that satisfy (1) become polynomial equations. Therefore the equilibrium constant K_{ij}, can be defined as

$$K_{ij} = |w_i^C w_j| \cdot (|w_i^C||w_j|)^{-1} = e^{-\frac{\Delta G_{ij}}{R \cdot T}}, \tag{2}$$

where $|w_i^C|$ and $|w_j|$ are the concentrations of the unhybridized (free) w_i^C and w_j respectively, ΔG_{ij} is the Gibbs free energy of the complex $w_i^C w_j$, R the Boltzman constant, and T the temperature in Kelvin. The Gibbs free energy for DNA complexes is a function of their sequence content and can be estimated using nearest neighbor (NN) thermodynamics parameters, which are available in the literature [21], [24], [36]. We should highlight that these constants depend on the reaction temperature and salt concentration; hence they have to be readjusted for each experiment. Also, the NN model assumes a two state (all or none) model of hybridization and has been developed for a single kind of duplex in solution. Although more complex and accurate models exist (multi-state models or next-to-NN models) [31], [37], when the probes are small in length and secondary intermediate structures are not expected, the two state model is a valid starting point [21]. Both conditions hold in our case since the probes are relatively small in length and secondary structures are not anticipated due to the careful design of the individual codewords that comprise the database elements and the query.

Let us assume a source encoded using N integers $j = 0, \ldots, N-1$, with known probabilities $P(j)$. Since an integer j is encoded into a codeword w_j, its probability is $P(w_j) = P(j)$. Furthermore, knowing that there are $N_T = M \cdot k$ occurrences of codewords in M database elements, the initial concentration $|w_j|_o$ is given by

$$|w_j|_o = P(w_j) \cdot M \cdot k \cdot C. \tag{3}$$

The mass conservation equations for the query and codeword are:

$$\left|w_i^C\right|_o = \left|w_i^C\right| + \sum_{j=0}^{N-1} \left|w_i^C w_j\right| = \left|w_i^C\right| + \sum_{j=0}^{N-1} K_{ij} \left|w_i^C\right| \cdot |w_j|, \tag{4}$$

$$|w_j| = \frac{|w_j|_o}{1 + K_{ij} |w_i^C|}. \tag{5}$$

The above system can be solved by substituting (5) into (4), to obtain

$$\left|w_i^C\right|_o = \left|w_i^C\right| + \sum_{j=0}^{N-1} K_{ij} \left|w_i^C\right| \cdot \frac{|w_j|_o}{1 + K_{ij} |w_i^C|}. \tag{6}$$

Following the steps in [16], [38], it can be shown that due to its monotonicity (6) has (i) a unique solution, denoted as $\left|w_i^C\right|_B$, that can be found using the bi-section method and (ii) has the following approximate solution that can be found under the assumption that the query concentration is smaller than the concentration of database elements ($\rho = |w_i^C|_o / C < 1$)

$$\left|w_i^C\right| \approx \frac{\left|w_i^C\right|_o}{M \cdot k \cdot C \cdot \hat{K}}, \tag{7}$$

where $\sum_{j=1}^{N} K_{ij} \cdot P(w_j) = \hat{K}$.

Using now (2) and (5) the concentration of each complex is

$$\left|w_i^C w_j\right| = K_{ij} \cdot \left|w_i^C\right| \cdot |w_j| = \frac{|w_j|_o \cdot K_{ij} \cdot \left|w_i^C\right|}{1 + K_{ij} \left|w_i^C\right|}. \quad (8)$$

Finally, by substituting $\left|w_i^C\right|$ from (7) we obtain

$$\left|w_i^C w_j\right| = \frac{M \cdot k \cdot C \cdot P(w_j) \cdot K_{ij} \cdot \left|w_i^C\right|_o}{M \cdot k \cdot C \cdot \hat{K} + K_{ij} \left|w_i^C\right|_o}. \quad (9)$$

C. Query Selectivity

Query selectivity SA_{ij} for a complex $w_i^C w_j$ can be defined as the percentage of its concentration within all the hybridized complexes,

$$SA_{ij} = \frac{\left|w_i^C w_j\right|}{\sum_{r=0}^{N-1} |w_i^C w_r|} = \frac{\left|w_i^C w_j\right|}{|w_i^C|_o - |w_i^C|}, \quad (10)$$

where use of (4) was made in obtaining the second equality. Query selectivity is a dimensionless quantity that can be seen as the probability of the complex $w_i^C w_j$ occurring in an ensemble of other complexes. Substituting (7) in the above equation we have

$$SA_{ij} = \frac{\left|w_i^C w_j\right|}{|w_i^C|_o - \frac{\left|w_i^C\right|_o}{M \cdot k \cdot C \cdot \hat{K}}}. \quad (11)$$

It is very common in the analysis of concentrations of molecular systems to evaluate ratios of concentrations or ratios of selectivities. This is very useful, for example, when examining the ratio of a desired hybridization (event) to an undesired one. In our case the selectivity ratios can be defined as:

$$\frac{SA_{ij}}{SA_{ij'}} = \frac{\left|w_i^C w_j\right|}{\left|w_i^C w_{j'}\right|} = \frac{K_{ij}}{K_{ij'}} \cdot \frac{|w_j|_o}{|w_{j'}|_o} \cdot \frac{1 + K_{ij'} \cdot \left|w_i^C\right|}{1 + K_{ij} \cdot \left|w_i^C\right|} \quad (12)$$

Utilizing the approximate solution of (7) and (3) we get

$$\frac{\left|w_i^C w_j\right|}{\left|w_i^C w_{j'}\right|} = \frac{K_{ij}}{K_{ij'}} \cdot \frac{P(w_j)}{P(w_{j'})} \cdot \frac{M \cdot k \cdot \frac{C}{\left|w_i^C\right|_o} \cdot \hat{K} + K_{ij'}}{M \cdot k \cdot \frac{C}{\left|w_i^C\right|_o} \cdot \hat{K} + K_{ij}}. \quad (13)$$

Equation (13) illustrates that at dilute concentrations the ratio of concentrations of two complexes is analogous to the ratio of their equilibrium constants (which is expected), but it is also analogous to a second term that highlights the dependency on the ensemble of fragments. A hint about this dependency is given in [18] when experimental findings are discussed. In [16] we derived the dependency term, which is a more generic version of (13). In [34] a similar result is presented following a quadratic approximation. Also in [39] a same conclusion is reached on the effect of concentration of undesired hybrids, following a kinetic analysis and considering simplified scenarios (eg., two probes attached on surface and two targets).

D. Signal-To-Noise Ratio

Similarly to [34] we can define the Signal-to-Noise Ratio (SNR) of a search with query w_i^C as:

$$SNR(w_i^C) = \frac{\sum_{j \in \mathbb{D}} \left|w_i^C w_j\right|}{\sum_{j \notin \mathbb{D}} |w_i^C w_j|}, \quad (14)$$

where $\mathbb{D}$ denotes the set of desired reactions. In our case desired hybridizations $w_i^C w_j$ are those for which the MSE of their corresponding signal values is less than or equal to the parameter T_P, while un-desired hybridizations are all the rest. However, since we only have codeword pair interactions and they satisfy NTC (see Section II and Appendix), desired and un-desired hybridizations can be specified and quantified. According to the NTC, desired complexes are those for which $|i - j| \leq T_P$, while un-desired are those for which $|i - j| > T_P$. Hence, (14) becomes

$$\begin{aligned} SNR(w_i^C) &= \frac{\sum_{j=i-T_P}^{i+T_P} \left|w_i^C w_j\right|}{\sum_{j=0}^{i-T_P-1} \left|w_i^C w_j\right| + \sum_{j=i+T_P+1}^{N-1} \left|w_i^C w_j\right|} \\ &= \frac{1 + \sum_{j=i-T_P}^{i-1} \frac{\left|w_i^C w_j\right|}{\left|w_i^C w_i\right|} + \sum_{j=i+1}^{i+T_P} \frac{\left|w_i^C w_j\right|}{\left|w_i^C w_i\right|}}{\sum_{j=0}^{i-T_P-1} \frac{\left|w_i^C w_j\right|}{\left|w_i^C w_i\right|} + \sum_{j=i+T_P+1}^{N-1} \frac{\left|w_i^C w_j\right|}{\left|w_i^C w_i\right|}}, \end{aligned} \quad (15)$$

where the righthand part was derived by dividing the nominator and denominator by $\left|w_i^C w_i\right|$. The value of $SNR(w_i^C)$ can be calculated by substituting (12) into the above equation.

IV. Parallel Query Retrieval

We consider now the case when N_Q queries $Q_1, Q_2, ..., Q_{N_Q}$ are utilized in retrieving information from the database. We have to consider the answer to two types of questions: (I) do any of the queries $Q_\imath$ hybridize to desired targets in the database? or (II) which of the queries $Q_\imath$ hybridize successfully in the database? Case (I) can be easily implemented in parallel without altering the retrieval protocol, by labeling each query with the same fluorescent dye. Case (II), however, can either be implemented in parallel or in series. Parallel implementation involves the fluorescent labeling with N_Q different dyes or with DNA microarrays where the location of each spot identifies the query. (DNA microarrays are small, solid supports onto which thousands of DNA sequences (probes) are immobilized at fixed locations forming an arrangement of spots.) Serial implementation involves iterative repetitions of single query searches that we described in the previous sections.

In this section we examine the parallel implementation of Cases I and II. Their analysis is identical up to the point of performance evaluation. We will follow the approach of Section III, for estimating $|Q_\imath F_i^\jmath|$ with the following two differences: (i) we now have N_Q queries, $Q_1, Q_2, ..., Q_{N_Q}$, with initial concentrations $|Q_1|_o, |Q_2|_o, ..., \left|Q_{N_Q}\right|_o$, respectively; (ii) the restriction that the query is perfectly aligned with the words in a database element is removed, thus making the analysis more general. In this case we can write the following reactions for each $Q_\imath$, $\imath = 1, ..., N_Q$:

$$\begin{aligned} Q_1 + F_i^\jmath &\underset{K_r}{\overset{K_f}{\rightleftharpoons}} Q_1 F_i^\jmath, \quad \forall i, \jmath \\ &\vdots \\ Q_\imath + F_i^\jmath &\underset{K_r}{\overset{K_f}{\rightleftharpoons}} Q_\imath F_i^\jmath, \quad \forall i, \jmath \\ &\vdots \\ Q_{N_Q} + F_i^\jmath &\underset{K_r}{\overset{K_f}{\rightleftharpoons}} Q_{N_Q} F_i^\jmath, \quad \forall i, \jmath \end{aligned} \quad (16)$$

As before, for each reaction there is an equilibrium constant. We will augment our notation here to accommodate the query index $\imath$. Therefore $K_{i,\imath}^\jmath$ is the equilibrium constant of the complex $Q_\imath F_i^\jmath$ and is related to the corresponding Gibbs free energy $\Delta G_{i,\imath}^\jmath$ by:

$$K_{i,\imath}^\jmath = \frac{|Q_\imath F_i^\jmath|}{|Q_\imath| \cdot |F_i^\jmath|} = e^{-\frac{\Delta G_{i,\imath}^\jmath}{R \cdot T}}. \quad (17)$$

Using this equation, the conservation equation for each query is

$$|Q_\imath|_o = |Q_\imath| + |Q_\imath| \sum_{i,\jmath} K_{i,\imath}^\jmath \, |F_i^\jmath|, \forall \imath, \tag{18}$$

where again the sum is over N_T terms, which is the number of complexes in the system. Subsequently, we can define $\alpha_\imath$ as

$$\begin{aligned} \alpha_\imath = \frac{|Q_\imath|_o - |Q_\imath|}{|Q_\imath|_o} &= \frac{\sum_{i,\jmath} |Q_\imath F_i^\jmath|}{|Q_\imath| + \sum_{i,\jmath} |Q_\imath F_i^\jmath|} \\ &= \frac{\sum_{i,\jmath} K_{i,\imath}^\jmath \, |F_i^\jmath|}{1 + \sum_{i,\jmath} K_{i,\imath}^\jmath \, |F_i^\jmath|}, \end{aligned} \tag{19}$$

where we used (18) in the last part. If we rewrite $\alpha_\imath$ as $\alpha_\imath = 1 - q_\imath$, where $q_\imath = \frac{|Q_\imath|}{|Q_\imath|_o}$, then (19) becomes

$$q_\imath = \frac{1}{1 + \sum_{i,\jmath} K_{i,\imath}^\jmath \, |F_i^\jmath|}. \tag{20}$$

Utilizing the conservation equations for each species we have

$$|F_i^\jmath| + \sum_{\imath'=1}^{N_Q} K_{i,\imath'}^\jmath \, |Q_{\imath'}| \, |F_i^\jmath| = |F_i^\jmath|_o \quad \forall i, \jmath. \tag{21}$$

By substituting (19) into (21) and solving for $|F_i^\jmath|$ we obtain

$$|F_i^\jmath| = \frac{|F_i^\jmath|_o}{1 + \sum\limits_{\imath'=1}^{N_Q} K_{i,\imath'}^\jmath \, |Q_{\imath'}|_o \, (q_{\imath'})} \tag{22}$$

The $N_T \times N_Q$ system of (20) and (22) can be solved iteratively for $q_\imath$ and $|F_i^\jmath|$, by assuming an initial value for each $q_\imath$ and alternating the use of the two equations. This procedure finds a single solution for $q_\imath$ and $|F_i^\jmath|$ regardless of the initial guess of $q_\imath$. With known $q_\imath$ and $|F_i^\jmath|$ now the sought after $|Q_\imath F_i^\jmath|$ can be found using the fact that $q_\imath = \frac{|Q_\imath|}{|Q_\imath|_o}$ and (17). We should note that by using the normalized query concentration, $q_\imath$, we avoid numerical errors that might have occurred if we were to solve for $Q_\imath$ directly [35].

Applying the assumptions of Section III, equations (20) and (22) assume a simpler form. In this case $F_i^\jmath = w_\jmath$, $K_{i,\imath'}^\jmath = K_{\imath,\jmath}$ with initial concentration as in (3), and $Q_\imath \in [w_1^C, ..., w_N^C]$ with initial concentration $|w_i^C|_o$. (Clearly for single word queries $N_Q \leq N$.) Equations (20) and (22) can be rewritten as

$$q_\imath = \frac{1}{1 + \sum\limits_{\jmath=1}^{N} K_{\imath\jmath} \cdot |w_\jmath|}, \tag{23}$$

$$|w_\jmath| = \frac{|w_\jmath|_o}{1 + \sum\limits_{\imath'=1}^{N_Q} K_{\imath'\jmath} \, |w_i^C|_o \, (q_\imath)}. \tag{24}$$

A. Retrieval Efficiency of Parallel Querying Scenarios

The retrieval error of the whole system can be used to study its performance. For each case, two definitions can be given depending on the laboratory protocol used to track and retrieve each query.

For Case I the overall error can be defined as the ratio of undesirable events (noise) over all events (signal+noise); that is,

$$E_I = \frac{\sum\limits_{\imath=1}^{N_Q} \sum\limits_{i,\jmath \notin \mathbb{D}} |Q_\imath F_i^\jmath|}{\sum\limits_{\imath=1}^{N_Q} \sum\limits_{i,\jmath} |Q_\imath F_i^\jmath|} \tag{25}$$

Note that the error of each individual query for Case I cannot be specified, since the fluorescent response of each query can not be distinguished. There are two sources of undesirable events for a query $Q_\imath$: (a) the fluorescent response from un-desired hybridizations between $Q_\imath$ and the database, and (b) the fluorescent response from all hybridizations between queries $Q_{\imath'}$, $\imath' \neq \imath$ and the database. Then the numerator of $\epsilon_\imath$ is the sum of the two noise sources that is $\sum\limits_{i,\jmath \notin \mathbb{D}} |Q_\imath F_i^\jmath| + \sum\limits_{i,\jmath,\imath' \neq \imath} |Q_{\imath'} F_i^\jmath|$ while the denominator is the same as in (25). However, we can see that $\sum\limits_{\imath=1}^{N_Q} \epsilon_\imath \neq E_I$, which would have been expected from such a system.

For Case II the overall error is:

$$E_{II} = \prod_{\imath'=1}^{N_Q} \epsilon_{\imath'}, \tag{26}$$

where $\epsilon_{\imath'} = \sum\limits_{i,\jmath \notin \mathbb{D}} |Q_{\imath'} F_i^\jmath| \; / \; \sum\limits_{i,\jmath} |Q_{\imath'} F_i^\jmath|$, is the error of each query. With single word queries and only word-to-word interactions then

$$\epsilon_{\imath'} = (1 + SNR(w_{\imath'}))^{-1}, \tag{27}$$

where $SNR(w_{\imath'})$ is given by (14).

We can define the SNR of the system as:

$$SNR = 1/E - 1, \tag{28}$$

where E is either E_I or E_{II}. For Case II the SNR for each individual query can also be defined by

$$SNR_\imath = 1/\epsilon_\imath - 1. \tag{29}$$

B. Estimating Source Statistics of a DNA Database

In this section we are considering the problem of estimating the source statistics of a DNA database. Although we synthesized the database from digital data and hence the initial source distribution was known, this distribution might have been altered due to various factors, such as synthesis and replication errors, dilution procedures, and ligation errors [11]. There exist common laboratory protocols (DNA microarrays) that can be used to estimate these statistics, or equivalently estimate the relative concentration of each word in the database. Alternatively, quantitative real time PCR (qPCR) can be used [40]. qPCR relies on probe target hybridizations that take place in solution and does not suffer from the complications of surface interfaces and other limitations that DNA microarrays introduce [41]. However, qPCR can be time consuming due to its serial fashion, since each probe has to be tested separately. Recently, parallel quantitative PCR has been developed to improve the throughput of qPCR [42] and therefore can be considered as an excellent candidate for our problem. We will present our analysis based on a microarray; however, we will use a more simplified hybridization model that does not account for the surface's effect on molecular kinetics, thus making our model directly applicable to a qPCR scenario.

Let us assume a DNA database formed by concatenating words w_j. Our goal is to find $|w_j|_o$. We assume that the database elements have a structure that allows PCR replication. Database elements are encapsulated by two known sequences (called primers) that are common to all elements, see [11] for more details. A sample of the database is taken and a PCR step with the two primers incorporates fluorescently labeled nucleotides. This procedure generates database elements that are fluorescently labeled and can be tracked.

Further, let us assume the existence of a small microarray with each spot containing as probes, all N Watson Crick complement words w_i^C, each with equal concentration C. To improve the accuracy, probe repetition among different spots on the array can also be used, but does not affect our analysis below. In a trivial microarray experiment a sample of the fluorescently labeled database is deposited on the microarray. Probes on the spots and words in the database react. The

final outcome is a microarray image with intensity $I_\imath$ at each spot. The following relation holds for $I_\imath$ and $|w_i^C|$:

$$I_\imath = f\left(\frac{\left|w_i^C\right|_{\text{bound}}}{\left|w_i^C\right|_o}\right) + \beta_\imath, \tag{30}$$

where $f()$ is a function relating concentration and intensity, and $\beta_\imath$ is the noise associated with the scanning apparatus [43]. Note that modeling hybridization reactions in solid-solution phase systems, as in microarrays, is a topic of great interest, see for example [44]–[46]. In [44] it has been shown that when equilibrium is reached, the model can be reduced to $I_\imath \propto \left|w_i^C\right|_{\text{bound}} / \left|w_i^C\right|_o$ if the instrument noise is ignored. The previous model is also used in [47] to estimate total concentrations in equilibrium. However, other parameters, such as the presence of the microarray surface, may affect the rate at which equilibrium is achieved [44] but we will ignore these parameters since we are not interested in the speed of the process we can assume that adequate incubation time has been allowed to reach equilibrium. Alternatively, qPCR which is a faster and solution phase only assay, can be used in lieu of microarrays or in tandem.

The above experiment can be formulated as a parallel multi-query search as described previously in Section IV. In this case $Q_\imath \equiv w_i^C$ and $N_Q = N$. Therefore,

$$\left|w_i^C\right|_o = \left|w_i^C\right| + \left|w_i^C\right| \sum_{j=0}^{N-1} K_{\imath,j} \left|w_j\right|, \quad \forall \imath. \tag{31}$$

From (19) we obtain

$$\alpha_\imath = \frac{\left|w_i^C\right|_o - \left|w_i^C\right|}{\left|w_i^C\right|_o} = \frac{\left|w_i^C\right|_{\text{bound}}}{\left|w_i^C\right|_o}. \tag{32}$$

From the above we see that $a_\imath = I_\imath$, then using (32) (31) becomes

$$\frac{I_\imath}{1 - I_\imath} = \sum_{j=0}^{N-1} K_{\imath,j} \left|w_j\right| \quad \forall \imath. \tag{33}$$

The above linear system of equations can be solved for $|w_j|$

$$\overline{W} = \mathbf{K}^{-1} \cdot \overline{I_Q}, \tag{34}$$

where $\overline{W} = [|w_1|, \ldots, |w_N|]^T, \overline{I_Q} = [\frac{I_1}{1-I_1}, \ldots, \frac{I_N}{1-I_N}]^T$, and $\mathbf{K}$ is an $N \times N$ matrix of all equilibrium constants.

With the estimated $|w_j|$, $|w_j|_o$ can be found from (24) as

$$\left|w_j\right|_o = \left|w_j\right| \cdot \left(1 + \sum_{\imath'=1}^{N_Q} K_{\imath' \jmath} \left|w_i^C\right|_o (1 - \alpha_\imath)\right). \tag{35}$$

V. Retrieval Efficiency of an Infinitely Large Database

In this section we will derive expressions for the SNR and the retrieval error of the system as the number of database elements reaches infinity. We will prove that a system that allows for a noise tolerant retrieval (as in NTC) has a lower retrieval error than other (traditional) systems without noise tolerance. A condensed form of the analysis analysis of this section also appeared in [38].

The study of retrieval error (efficiency, accuracy) is of critical importance in designing memory systems. In [5] it was shown that the information capacity increases exponentially with the size of the index; however, it was claimed that the protocol has low retrieval error due to the use of nested PCR (another form of PCR). In [48] a formula to describe the channel capacity of molecular machines was presented. The analysis is based on the Brownian motion of molecules and the maximum possible information gain. This gain is a function of the energy that a molecular machine dissipates into the surrounding medium, the thermal noise energy which disturbs the machine, and the number of independently moving parts involved in the operation. Simulations were used in [49] to quantify retrieval efficiency and similar conclusions as the ones derived mathematically in this section were drawn. Our analysis also reaches similar conclusions as in [27], [34] but from another point of view.

As in Section III we will assume that only perfect aligned complexes are present. When more database elements are introduced into the database ($M \to \infty$) the number of codewords increases and therefore their concentration increases, that is $\lim_{M\to\infty} |w_j|_o \to \infty$, while the concentration of the queries is bounded (query in dilute). From (13) after some basic steps we obtain

$$\lim_{M\to\infty} \frac{\left|w_i^C w_j\right|}{\left|w_i^C w_i\right|} = \frac{\infty}{\infty} = \ldots = \frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}. \tag{36}$$

From (15) and using the previous equation we obtain

$$\begin{aligned}
&\lim_{M\to\infty} SNR(w_i^C) = SNR(w_i^C)_\infty = \\
&= \frac{1 + \sum_{j=i-T_P}^{i-1} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right) + \sum_{j=i+1}^{i+T_P} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right)}{\sum_{j=0}^{i-T_P-1} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right) + \sum_{j=i+T_P+1}^{N-1} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right)} \\
&= \frac{\sum_{j=i-T_P}^{i+T_P} (K_{ij} \cdot P(w_j))}{\sum_{j=0}^{i-T_P-1} (K_{ij} \cdot P(w_j)) + \sum_{j=i+T_P+1}^{N-1} (K_{ij} \cdot P(w_j))}.
\end{aligned} \tag{37}$$

Note that according to the NTC in addition to w_i, codewords neighboring w_i are retrieved, and hence the codewords are designed such that $\frac{K_{ij}}{K_{ii}} < 1$, for $|i-j| \le T_P$ (numerator of (37)). On the other hand for codewords outside this neighborhood (denominator of (37)) $\frac{K_{ij}}{K_{ii}} \ll 1$, for $|i-j| > T_P$.

We can calculate the corresponding retrieval error as

$$E_\infty(w_i^C) = \lim_{M\to\infty} E(w_i^C) = \frac{1}{1 + SNR(w_i^C)_\infty}. \tag{38}$$

If we assume a uniform distribution for $P(w_j)$ we obtain

$$\lim_{M\to\infty} SNR(w_i^C) = \frac{\sum_{j=i-T_P}^{i+T_P} K_{ij}}{\sum_{j=0}^{i-T_P-1} K_{ij} + \sum_{j=i+T_P+1}^{N-1} K_{ij}}. \tag{39}$$

Then the corresponding retrieval error is in agreement with the definition of *computational incoherence* in [27], [34] (the probability of error); However, the approach in [27], [34] is rather qualitative than quantitative. In this section we showed (using a linear approximation), that at infinity the retrieval error is only a function of the source statistics $P(w_j)$ and the equilibrium constants K_{ij}.

We can further compare the performance of the proposed retrieval and codeword design system with codeword designs that allow only exact matching (perfect hybridization). Assume that we have such a codeword set K'_{ij}, we can find the SNR of a search w'_i at infinity by replacing K_{ij} with K'_{ij}, and setting $T_P = 0$ in (39), that is,

$$\lim_{M\to\infty} SNR(w_i'^C) = \frac{K'_{ii}}{\sum_{j=0}^{i-1} K'_{ij} + \sum_{j=i+1}^{N-1} K'_{ij}}. \tag{40}$$

By comparing (40) and (39) we see that for the SNR expression in (40) to be larger than or equal to the expression in (39) either

$$K'_{ii} \ge \sum_{j=i-T_P}^{i+T_P} K_{ij} \tag{41}$$

or

$$\sum_{j=0}^{i-1} K'_{ij} + \sum_{j=i+1}^{N-1} K'_{ij} \leq \sum_{j=0}^{i-T_P-1} K_{ij} + \sum_{j=i+T_P+1}^{N-1} K_{ij}. \quad (42)$$

We have shown that controlled cross-hybridization is actually beneficial in terms of SNR when designing such systems, since (41) or (42) need to be satisfied by a system allowing only perfect hybridization. The same conclusion was reached in [50] when the performance of microarray systems was evaluated using a communication model.

VI. Computer Simulation of DNA Databases

In this section we present the results obtained when the models and derivations of the previous section are implemented in a computing language to simulate data retrieval in a test DNA database. The simulation language used was MATLAB. We used the codeword set shown in Appendix to encode $N = 32$ signal values. We present our results on the relationship between annealing selectivities and source statistics. Finally, we show our numerical findings on the performance of infinitely large databases. For all experiments the initial concentration of each database element was $C = 10^{-5} {}^{mol}/_{Liter}$.

A. Results on Annealing Selectivities and Source Statistics

We model each word's w_j probability as $P(w_j) = P(j) + \epsilon_j$, where $P(j)$ is the probability of the index j and ϵ_j is a random variable that follows a uniform distribution with mean $E[\epsilon_j] = 0$ and standard deviation σ_ϵ. This scenario will simulate cases when the source statistics are different from the ones initially assumed. In fact, it will be shown that the effect on the resulting SNR is less than an order of magnitude.

Our experimental setup was a database with $M = 20$ database elements and $k = 20$ words per database element. The reaction temperature was $T = 60^oC$. We found the Gibbs free energy of all possible pairs (32^2) and their corresponding equilibrium constants using the methodology of Section III and (2). We performed our simulations for ρ equal to $100, 10, 1, 0.1$, and 0.01. We also assumed a uniform distribution for the source, that is $P(j) = 1/32$. We tested for two different standard deviations $\sigma_\epsilon = \{1.86 \cdot 10^{-3}, 4.65 \cdot 10^{-3}\}$ which correspond to the two ranges for ϵ_j, $[\frac{-0.2}{32}, \frac{0.2}{32}]$ and $[\frac{-0.5}{32}, \frac{0.5}{32}]$. For each ρ we generated 50 sets of random variables ϵ_j with the above standard deviations. Afterwards, we found the SNR of a query search with $q_d = \{15\}$ against the database for a given ρ and σ_ϵ. In Fig. 2 we show a box-whisker plot for each σ_ϵ for various ρ. (Different box size indicates the case of σ_ϵ.) With the marker '$\times$' we denote the SNR of a database where $P(w_j) = P(j)$, for comparison. We see that the median for each case (illustrated by a line inside the box) is close to the actual SNR marked with '$\times$'. In fact the effect of deviation from uniformity on the SNR is minimal.

In a similar fashion, we wanted to test the effect of different source distributions on the performance of query retrieval. We assume again that we are querying with $q_d = \{15\}$ and that $\rho = 10$. Now we assume that the source follows a Gaussian distribution with mean $m = j$, where j is a word index, and variance σ^2. In the following experiment we compared all indices $[0, \ldots, 31]$ vs. $\sigma^2 = [3, 4, 5]$, for $T_P = 3$. In Fig. 3 we plot in logarithmic scale the corresponding SNR values. For example we see that when $m = 4, \sigma^2 = 3$ the SNR is close to 100, despite the fact that the Gaussian is biased towards codeword 4. We would expect the SNR values to be maximized for $m = 15$ which corresponds to our query q_d. However, although it is not clear from the graph, the maximum SNR is achieved for $m = 14$. After examining the formula of the SNR for the two cases $m = 14$ and $m = 15$, we found that although the numerators are the same, the denominator for $m = 15$ is larger than the one for $m = 14$ and hence the smaller SNR for $m = 15$. This is attributed to the codeword design set and highlights the importance of using such simulations and models when evaluating codeword sets.

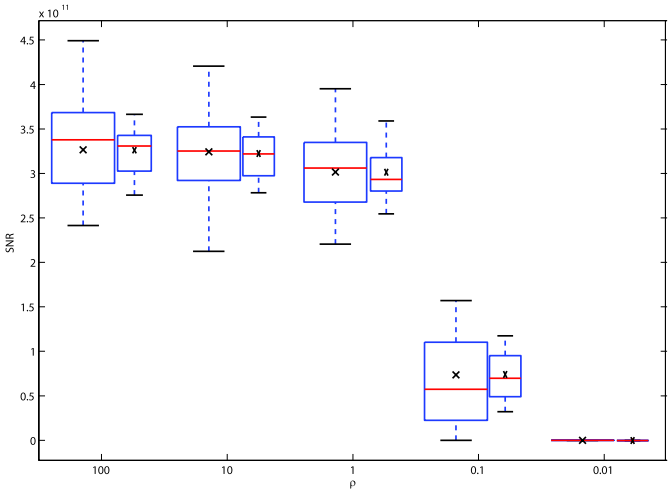

Fig. 2. SNR values for a database with uniformly varying source statistics with standard deviation $\sigma_\epsilon = 1.86 \cdot 10^{-3}$ and $\sigma_\epsilon = 4.65 \cdot 10^{-3}$ shown in the large and small box-whisker respectively. ('$\times$' denotes the SNR of a database with $P(w_j) = P(j)$)

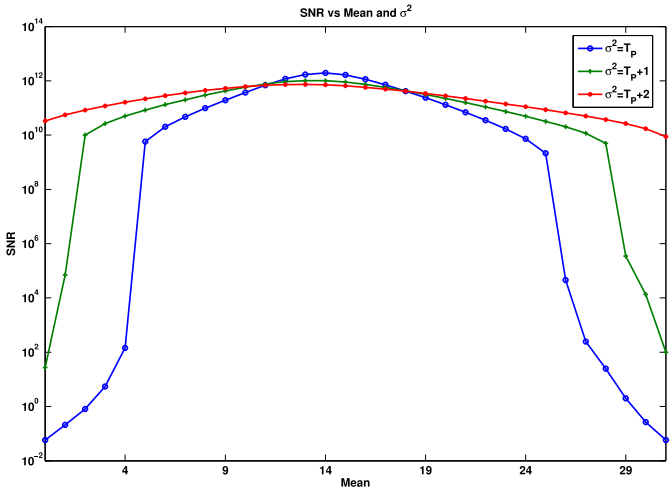

Fig. 3. SNR values for a database with Gaussian varying source statistics with means centered at each codeword index and variance T_P, T_P+1, T_P+2.

From our simulation experiments we conclude that in order to have a retrieval system that is not biased by the source statistics the source should follow a uniform distribution. Although, this might be hard to enforce in single word queries, for multiple word queries this requirement will be easier to satisfy.

B. Retrieval Results of an Infinitely Large Database

We first compared the accuracy of the approximate solution of (6) provided by (7) with the exact numerical solution. As a comparison metric we chose the SNR. We used the same words as before. Our query was the integer $q_d = \{14\}$. We found the equilibrium constants between $q_d = \{14\}$ and the signal values $0, \ldots, 31$. We assumed a uniform distribution for the source, that is $P(w_j) = 1/32, j = 0, \ldots, 31$. We used $M = 3$ database elements and $k = 20$ words per database element. According to (15) and $T_P = 3$ the SNR can be found as $SNR(w_{14}^C) =$

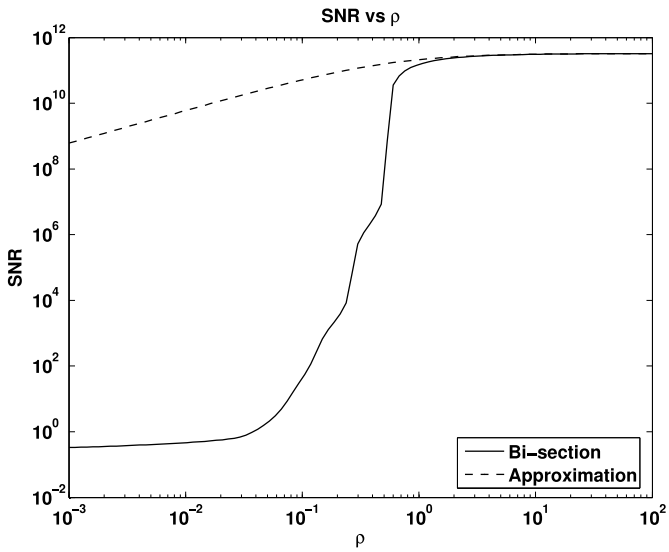

Fig. 4. SNR for various ρ.

$\sum_{j=11}^{17} \left|w_{14}^C w_j\right| \cdot (\sum_{j=0}^{10} \left|w_{14}^C w_j\right| + \sum_{j=18}^{31} \left|w_{14}^C w_j\right|)^{-1}$. We calculated the SNR for $\rho = C/\left|w_{14}^C\right|_o = 10^{-3}, \ldots, 10^2$ using the bisection method and the approximation. The results are shown in Fig. 4. Similarly, to the previous sections we can see that when $\rho > 1$ the approximation is very close to the exact solution. Furthermore, observe that the SNR increases as ρ increases, which is a clear indication that competitive hybridization is a critical and desired aspect for the performance of our system.

To find the SNR as the database size increases we repeated the above experiment but we increased the size of the database (M) at each iteration. Specifically, we want to verify the validity of (37). Actually we will see that $SNR(w_i^C)_M \leq SNR(w_i^C)_\infty$, that is (37) is an upper-bound for the SNR performance of a database with a finite number (M) of database elements.

In Fig. 5 the SNR of a database of size $M = 1, \ldots, 10^{10}$ for $\rho = 100, 10, 1, 0.1, 0.01$, is shown. The value of $SNR(w_{14}^C)_\infty = 3.2681 \cdot 10^{11}$, found using (39), is also shown with a dashed line. We see that for large M we achieve the bound at infinity independently of the value of ρ. Furthermore, as long as the query is in dilute, $\rho > 1$, the performance of the database is very close to the maximum achievable SNR. The graph also hints at an estimate of ρ relative to the database size. We see that if we chose $\rho > 1/M$ we can always have good performance. As a rule of thumb, $\rho = 1$ should be adequate to achieve good performance for databases with $M > 100$. Furthermore with basic curve fitting we can find a tighter upper bound for the SNR for any M and ρ as $SNR(w_i^C)_{M,\rho} \leq SNR(w_i^C)_\infty \cdot M/(M+\rho^{-1})$. Fig. 6 illustrates this relationship for the experimental setup we are considering in this section.

C. Simulation Results on Multiple Query Searches

To test the performance of the system with parallel queries we followed the analysis of Section IV. We used a similar set as above. Namely, we used the same 32 words of length 19 as in the previous section. Our queries were the integers 14, and 29. We found the equilibrium constants between 14, 29 and the signal values $0, \ldots, 31$. We assumed a uniform distribution, therefore $P(w_j) = 1/32$. We had M database elements and $k = 20$ words per database element. We also assumed equal query concentrations. We evaluated the SNR for the two cases SNR_I, SNR_{II} of Section IV-A for a database size of $M = 1, \ldots, 10^{10}$, and $\rho = 100, 10, 1, 0.1, 0.01$.

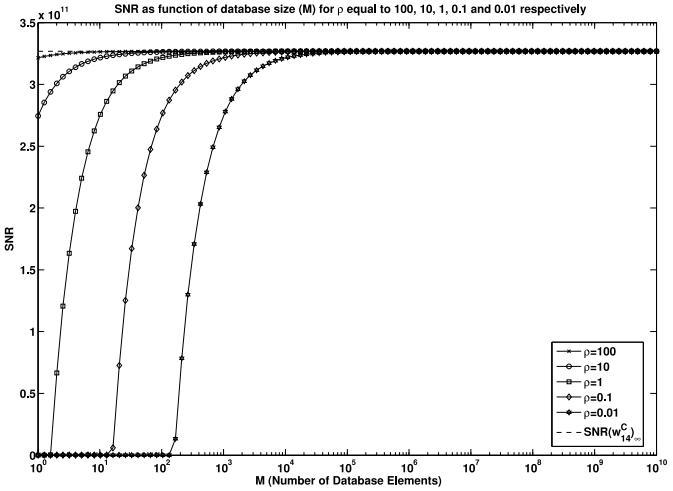

Fig. 5. SNR as a function of database size M, for various ρ. The upper bound $SNR(w_{14}^C)_\infty$ is plotted as a dashed line for comparison.

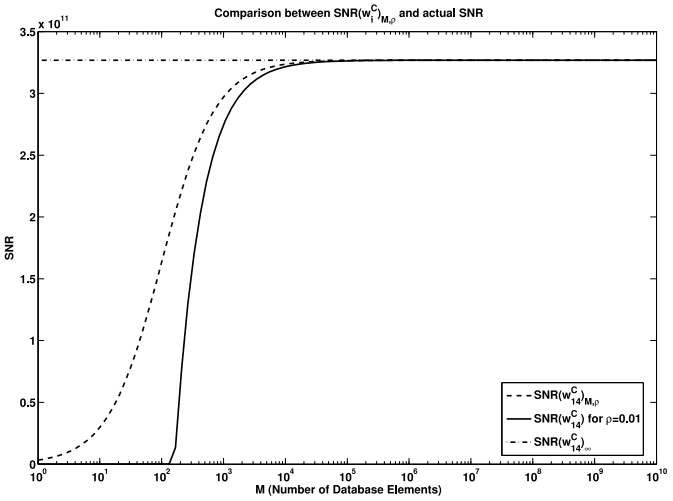

Fig. 6. Comparison between $SNR(w_i^C)_{M,\rho}$ of Eq. VI-B (dashed line) and the curve (solid line) for $\rho = 0.01$ of Fig. 5. The upper bound $SNR(w_{14}^C)_\infty$ is plotted as a dash-doted line for comparison

On the top of Fig. 7 we plotted the SNR for Case I where, an 'OR' type query against the two queries is performed. We see that for large ρ the performance of the database reaches quickly an upper bound when fewer than a $M = 100$ database elements are introduced. The upper bound of the SNR is equal to $1.749 \cdot 10^{10}$.

In the bottom of Fig. 7 we plotted the SNR for Case II where, an 'AND' type query against the two queries is performed. We see that for large ρ the performance of the database again reaches an upper bound with fewer than 100 database elements. The upper bound of the SNR is equal to $SNR_{II}^\infty = 2.938 \cdot 10^{21}$. Remarkably the bound SNR_{II}^∞ is the product of $SNR(w_{14}^C)_\infty$ and $SNR(w_{29}^C)_\infty$ which are equal to $3.268 \cdot 10^{11}$ and $8.99 \cdot 10^9$, respectively. We can argue therefore that the retrieval performance per query is independent of the number of simultaneous queries in the system, since the SNR_∞ for query 14 is the same for the single (see Section VI-B and Fig. 5) and the parallel query case (current section).

To verify the results of Section IV-B and to illustrate that the methodology for simulating multiple queries can be directly applied in microarray analysis we tested a case where we have a hypothetical microarray where each spot, from a total of 32 spots, contains as

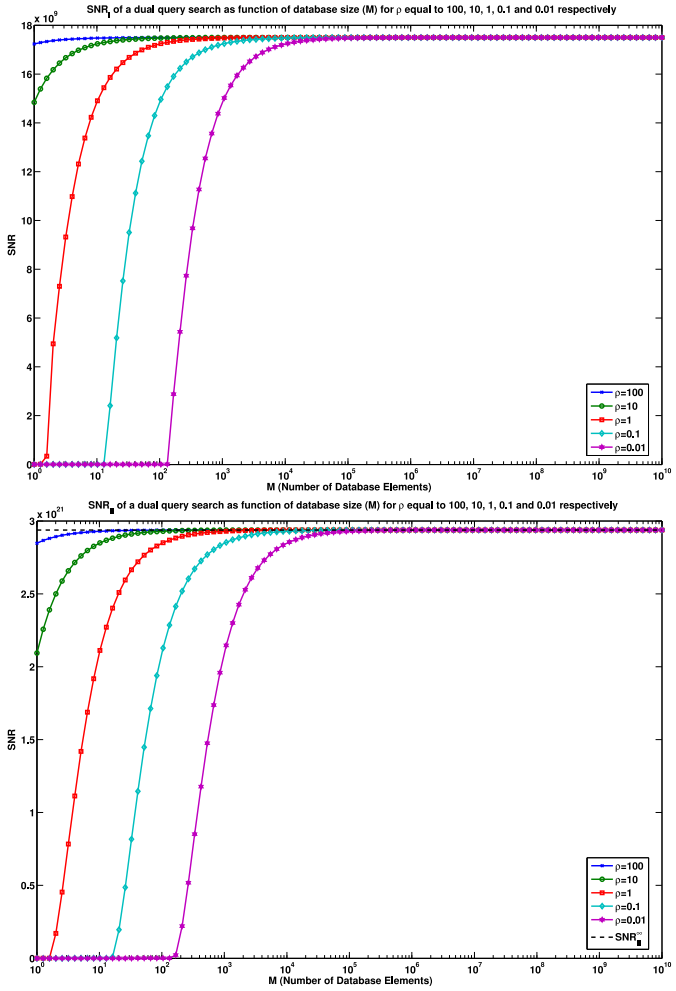

Fig. 7. SNR_I and SNR_{II} of a parallel two query search as a function of database size M, for various values of ρ. The upper bound SNR_{II}^{∞} is plotted as a dashed line for comparison on the bottom plot

probe w_i^C. We assumed that $|w_i^C|_o = 10^{-3}$ and that our database had $M = 100$ database elements with $k = 20$ words.

In order to find the spot intensities $I_\imath$ we assumed that the original statistics of the source and the concentration C of each database element were known. More specifically, we assumed a uniform distribution. Note that $|w_i^C|_o > C$ is required in order to simulate microarray reaction conditions. By iterating (23) and (24) we found the $I_\imath (= \alpha_\imath = 1 - q_\imath)$ and the annealing selectivities for the complexes $w_i^C w_j$.

For the case when we have no prior knowledge about the concentration of the words in the database, we work backwards, and estimate $|w_j|_o$ knowing only $|w_i^C|_o$, $I_\imath$, and the equilibrium constants $K_{\imath,j}$. We follow the analysis of Section IV-B, solve the system in (33), and compute $\widetilde{|w_j|_o}$ using (24). We found that the MSE between the actual values $|w_j|_o$ and the estimated values $\widetilde{|w_j|_o}$ was equal to $E[\left(|w_j|_o - \widetilde{|w_j|_o}\right)^2] = 5.95 \cdot 10^{-38}$, which can be attributed to numerical precision errors.

VII. Conclusion

In this paper we presented an elaborate framework to simulate single and parallel query scenarios. Our kinetic analysis and formulation allows for numerical solutions, as well as approximate solutions under certain conditions. When approximations are utilized, useful bounds on the performance of a DNA database can be derived. Specifically, we showed that the SNR of a DNA database is upper bounded by the SNR of an infinitely large DNA database that has the same source distribution. We also showed that microarray technology can be used to estimate the statistics of an 'unknown' DNA database. A number of simulation results were presented that verify and support our claims. Our simulations indicate that it is very critical to simulate codeword designs prior to experimentation in order to identify flaws in the design. We also found appropriate concentration of database and query in order to achieve good SNR in our retrieval. We also showed that the distribution of source is critical in the accuracy of the retrieval.

Our simulation framework has applications also in life sciences. It can be used to simulate and optimize laboratory protocols such PCR, primer and oligo design, microarray probe design and simulations. For example, SNR type metrics (section III-D) can be defined to assess the accuracy of hybridization of PCR primers or microarray probes. Although, there exists literature in the topic, the solution of coupled non-linear equations can be prohibitive computationally, especially when considering large probe target systems (e.g., microarrays contain more than 20,000 probes). The proposed linearization approach can provide a fast approximate solution that can be either used as a seed for more accurate computational solutions or provide an estimate of error, which may be adequate for some applications [35].

As far as future improvements are concerned, it is of great interest to find the speed of a query search. That is, how long do we have to wait until we get an answer from a DNA database? Is the search limited by the time it takes for the molecules to find each other (diffusion limited) or by the actual hybridization reaction and time to reach equilibrium (reaction limited) [51]? For example, it is known that microarray hybridization portrays diffusion limited characteristics [51]. To answer this type of questions, a kinetic and diffusion analysis is required. Such analysis requires the solution of coupled differential equations, a rather computationally intensive task. Furthermore, solving the differential equations requires the definition of the exact forward and reverse rates, which are system dependent and largely affected by environmental parameters (e.g., presence of surface, ionic conditions, viscosity of buffer). The solution of the system of differential equations yields time relaxation constants that portray the elapsed time necessary for the system to reach equilibrium and hence hint on the speed of a query search. It is clear therefore that real wet lab experiments are needed even in a small demonstrational scale to derive such parameters.

Appendix
Codeword Design

We design DNA codewords w_i, such that the hybridization strength between a codeword w_i and the Watson-Crick complement of another codeword w_j^C, as quantified by their melting temperature $T_M(w_i, w_j^C)$, is inversely proportional to the absolute difference of the corresponding encoded integer signal values $|i-j|$. To accomplish this, we introduced the Noise (or inexact match) Tolerance Constraint (NTC) [10], [52]:

$$\text{for} \quad w_i = \mathcal{C}(i) \quad \text{and} \quad w_j = \mathcal{C}(j)$$

$$T_M(w_i, w_j^C) = \begin{cases} \text{maximum} & if \quad i = j \\ \propto \frac{1}{f(|i-j|)} & if \quad |i-j| \leqslant T_P \\ < T & if \quad |i-j| > T_P \end{cases} \tag{43}$$

where $\mathcal{C}()$ is the mapping (or the look-up table), $f()$ represents a monotonically increasing function, T and T_P are user selected thresholds that control the noise tolerance of the set. This constraint

TABLE I
THE CODEWORDS OF LENGTH $l = 19$ FOR $N = 32$.

i	w_i	i	w_i
0	AAGGTCCAAAGTGCCACCC	16	AACGGCGTTCGTACCAGCC
1	AAGGTCCAAAGTACCGGCC	17	AAAGGCGTATGTACCGCCC
2	AAGGTCCAGAGAACCTGCC	18	AATGGCGAGCGTACCTTCC
3	AAGGTCTCTAGTTCCGGCC	19	AACTTCGCGTGTACCTTCC
4	AAGGTCAAGAGTACCGGCC	20	AACTGCGCTTGTACCTTCC
5	AAGGTCATAGGTTCCTGCC	21	AACTGCGAGTGAGCCTTCC
6	AAGGTCACAGGTTCCATCC	22	AATTGCGCGTGAGCCTTCC
7	AAGGTCAAGGGTCTCATCC	23	AATTGCTTGTGGTCCTTCC
8	AATTCCAAGGGTTCCATCC	24	AATTGCTTGTGATGCTTCC
9	AATATCAAGGGACCCATCC	25	AATTGCTCGTGTCGCTTCC
10	AATACCAAGGGATCCATCC	26	AATTGCTTATGGTGCTGCC
11	AAATCCGAGGGACCCATCC	27	AATTGCTCTAGTAGCCGCC
12	AATGGCGTGGGACTCATCC	28	AATTGCTCGAGTAGCCGCC
13	AACGGCGTAGGATCCATCC	29	AATTGCTAGAGTAGCCGCC
14	AAAGGCGTCGGTGCCATCC	30	AAATGCTCGAGTAGCCGCC
15	AACGGCGTTCGTGCCATCC	31	AAATCCACGAGTAGCCGCC

combined with other self and group constraints (originating from biochemical considerations), such as, the self-complementarity, consecutive bases, GC content, frame-shift, and the reverse complement constraints are needed to ensure that only wanted duplexes will be formed. In other words, the possibility of formation of unwanted duplexes for $|i - j| > T_P$ is minimized while the possibility of wanted ones for $|i-j| \leq T_P$ is maximized. In a laboratory setting this translates to minimizing the concentration of unwanted hybridizations while maximizing the concentration of wanted ones.

We have developed a number of stochastic algorithms that can generate codewords that satisfy the imposed constraints [29], [52]. Briefly explained, the algorithm presented in [29] starts with an initial solution set $\mathbf{S}$ arranged in a matrix form where each row represents a codeword. At each iteration, one of seven operators is selected at random or based on a schedule to create a new set $\mathbf{S}'$ of words. The available operators are: (i) randomly perturb all the columns, (ii) randomly interchange two columns, (iii) change randomly the bases of a randomly selected column, (iv) add a column of a randomly selected base at a random location, (v) remove a randomly selected column, (vi) add a column of random bases at a random location, and (vii) randomly change an element of $\mathbf{S}$.

The objective of the algorithm is to minimize a cost function $\mathfrak{C}$, which is defined as a weighted sum of all constraint violations, thus $min(\mathfrak{C}) = 0$. Therefore, if $\mathfrak{C}(\mathbf{S}') < \mathfrak{C}(\mathbf{S})$ then $\mathbf{S} = \mathbf{S}'$ and $i = 0$ (a stagnation counter); otherwise, $i = i + 1$. To allow the algorithm to escape from local minima we introduce a stochastic non-improving step: $\mathbf{S}'$ is accepted with probability $\vartheta = e^{\frac{d}{\alpha \cdot i}}$, where $d = \mathfrak{C}(\mathbf{S}) - \mathfrak{C}(\mathbf{S}')$, and α is a (fixed or updated over time) parameter. The algorithm terminates if the cost is zero and a solution to the constrained problem is returned or the maximum number of iterations has been reached and hence it returns a set that partially fulfils the constraints. A solution for 5-bit signals is tabulated in Table A, for $N = 32$, $l = 19$ and $T_P = 3$. All desired hybridizations have a melting temperature (the temperature above which a duplex is considered broken) above $T = 55.4^oC$.

ACKNOWLEDGMENT

The authors would like to thank Prof. Hatzimanikatis at EPFL, for useful discussions. They would also like to thank the reviewers for helpful suggestions. Finally, S.A. Tsaftaris would like to acknowledge the Onasis Public Benefit Foundation for the financial support while this work was being in development.

REFERENCES

[1] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, no. 11, pp. 1021–1024, November 1994.

[2] E. B. Baum, "Building an associative memory vastly larger than the brain," *Science*, vol. 268, no. 5210, pp. 583–585, April 1995.

[3] P. A. Morin, H. N. Poinar, G. Eglinton, O. A. Ryder, A. Mclaren, Y.-P. Zhang, S. Brenner, and K. Benirschke, "Preservation of DNA from endangered species," *Science*, vol. 289, no. 5480, pp. 725d–727, August 2000.

[4] A. Kameda, M. Yamamoto, H. Uejima, M. Hagiya, K. Sakamoto, and A. Ohuchi, "Conformational addressing using the hairpin structure of single-strand DNA," in *DNA Computing 9*, L. Wang, K. Chen, and Y. S. Ong, Eds., vol. 2943. Springer Berlin / Heidelberg, June 2004.

[5] S. Kashiwamura, M. Yamamoto, A. Kameda, T. Shiba, and A. Ohuchi, "Hierarchical DNA memory based on nested PCR," in *DNA Computing 8*, ser. Lecture Notes in Computer Science, M. Hagiya and A. Ohuchi, Eds., vol. 2568. Springer, 2002, pp. 112–123.

[6] J. H. Reif, T. H. Labean, M. Pirrung, V. S. Rana, B. Guo, C. Kingsford, and G. S. Wickham, "Experimental construction of very large scale DNA databases with associative search capability," *Lecture Notes in Computer Science*, vol. 2340, January 2002.

[7] M. Takinoue and A. Suyama, "Hairpin-DNA memory using molecular addressing," *Small*, vol. 2, no. 11, pp. 1244–1247, 2006.

[8] Y. Tsuboi, Z. Ibrahim, and O. Ono, "DNA computing approach to semantic knowledge representation," *Int. J. Hybrid Intell. Syst*, vol. 2, no. 1, pp. 1–12, 2005.

[9] J. Chen, R. Deaton, and Y.-Z. Wang, "A DNA-based memory with *in vitro* learning and associative recall," *Natural Computing*, vol. 4, no. 2, pp. 83–101, June 2005.

[10] S. A. Tsaftaris, A. K. Katsaggelos, T. N. Pappas, and E. T. Papoutsakis, "How can DNA computing be applied to digital signal processing?" *IEEE Signal Processing Mag.*, vol. 21, no. 6, pp. 57–61, 2004.

[11] S. A. Tsaftaris and A. K. Katsaggelos, "On designing DNA databases for the storage and retrieval of digital signals," *Lecture Notes in Computer Science*, vol. 3611, pp. 1192–1201, July 2005.

[12] S. A. Tsaftaris, V. Hatzimanikatis, and A. K. Katsaggelos, "DNA as a medium for storing digital signals," in *Proceedings of the 10th Int Conf on the Simulation and Synthesis of Living Systems*, vol. 1, June 2006, pp. 303–309.

[13] S. A. Tsaftaris, V. Hatzimanikatis, and A. K. Katsaggelos, "DNA hybridization as a similarity criterion for querying digital signals stored in DNA databases," in *Processings of ICASSP 2006*, vol. 2, May 2006, pp. II–1084–II–1087.

[14] S. A. Tsaftaris and A. K. Katsaggelos, "The not so digital future of digital signal processing," *Proceedings of the IEEE*, vol. 96, no. 3, pp. 375–377, 2008.

[15] R. M. Haralick and L. G. Shapiro, *Image Matching*, ser. Computer and Robot Vision. Reading, Massachusets: Addison-Wesley, 1993, vol. II, ch. 16.

[16] S. A. Tsaftaris, V. Hatzimanikatis, and A. K. Katsaggelos, "*In silico* estimation of annealing specificity of query searches in DNA databases," *Journal of the Japan Society of Simulation Technology (JSST) special issue "Application and Simulation of DNA Computing"*, vol. 24, no. 4, pp. 268–276, December 2005.

[17] D. C. Tulpan, H. H. Hoos, and A. E. Condon, "Stochastic local search algorithms for DNA word design," *Lecture Notes in Computer Science*, vol. 2568, pp. 229–241, 2003.

[18] D. Tulpan, M. Andronescu, S. B. Chang, M. R. Shortreed, A. Condon, H. H. Hoos, and L. M. Smith, "Thermodynamically based DNA strand design," *Nucleic Acids Research*, vol. 33, no. 15, pp. 4951–4964, 2005.

[19] M. Garzon, V. Phan, S. Roy, and A. Neel, "In search of optimal codes for DNAs computing," in *DNA Computing*, 2006, pp. 143–156.

[20] A. Condon, "Designed DNA molecules: principles and applications of molecular nanotechnology," *Nature Reviews Genetics*, vol. 7, no. 7, pp. 565–575, June 2006.

[21] J. Santalucia and D. Hicks, "The thermodynamics of DNA structural motifs," *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 415–440, 2004.

[22] G. L. Moore and C. D. Maranas, "Predicting out-of-sequence reassembly in DNA shuffling," *J Theor Biol*, vol. 219, no. 1, pp. 9–17, November 2002.

[23] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry: Part III: The Behavior of Biological Macromolecules (Their Biophysical Chemistry; PT. 3)*. W. H. Freeman, 1980.

[24] G. G. Hammes, *Thermodynamics and Kinetics for the Biological Sciences*. New York, NY: John Wiley & Sons, 2000.

[25] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3406–3415, July 2003.

[26] V. Ivanov, Y. Zeng, and G. Zocchi, "Statistical mechanics of base stacking and pairing in dna melting," *Physical Review E*, vol. 70, no. 5, pp. 051 907+, November 2004.
[27] J. A. Rose, R. J. Deaton, D. R. Franceschetti, M. Garzon, and S.E. Stevens Jr., "A statistical mechanical treatment of error in the annealing biostep of DNA computation," in *Proceedings of the Genetic and Evolutionary Computation Conference*, vol. 2, Orlando, Florida, USA, 13-17 July 1999, pp. 1829–1834.
[28] J.-Y. Wang and K. Drlica, "Modeling hybridization kinetics," *Mathematical Biosciences*, vol. 183, no. 1, pp. 37–47, May 2003.
[29] S. A. Tsaftaris and A. K. Katsaggelos, "A new codeword design algorithm for DNA based storage and retrieval of digital signals," in *Preproceedings of the 11th International Meeting on DNA-based computers DNA 11*, London, Ontario, Canada, 2005.
[30] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce, "Thermodynamic analysis of interacting nucleic acid strands," *SIAM Review*, vol. 49, no. 1, pp. 65–88, 2007.
[31] D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt, and D. H. Turner, "Predicting oligonucleotide affinity to nucleic acid targets," *RNA*, vol. 5, no. 11, pp. 1458–1469, November 1999.
[32] M. Garzon and C. Oehmen, "Biomolecular computation in virtual test tubes," in *DNA Computing*, 2002, pp. 117–128.
[33] M. Garzon, D. Blain, and A. Neel, "Virtual test tubes: For biomolecule-based computing," *Natural Computing*, vol. 3, no. 4, pp. 461–477, December 2004.
[34] J. A. Rose, R. J. Deaton, M. Hagiya, and A. Suyama, "Coupled equilibrium model of hybridization error for the DNA microarray and tag-antitag systems," *IEEE Trans. Nanobiosci.*, vol. 6, no. 1, pp. 18–27, 2007.
[35] M. T. Horne, D. J. Fish, and A. S. Benight, "Statistical thermodynamics and kinetics of DNA multiplex hybridization reactions," *Biophys. J.*, vol. 91, no. 11, pp. 4133–4153, December 2006.
[36] J. Santalucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc Natl Acad Sci U S A*, vol. 95, no. 4, pp. 1460–1465, February 1998.
[37] D. Jost and R. Everaers, "A unified Poland-Scheraga model of oligo- and polynucleotide DNA melting: Salt effects and predictive power," *Biophys. J*, vol. 96, no. 3, pp. 1056–1067, 2009.
[38] S. A. Tsaftaris and A. K. Katsaggelos, "Retrieval accuracy of very large DNA-based databases of digital signals," in *Proc. of 2007 European Signal Processing Conference, Poznań, Polland*, September 3-7 2007, pp. 1561–1567.
[39] Y. Zhang, D. A. Hammer, and D. J. Graves, "Competitive hybridization kinetics reveals unexpected behavior patterns." *Biophys. J.*, vol. 89, no. 5, pp. 2950–2959, November 2005.
[40] R. Higuchi, G. Dollinger, P. S. Walsh, and R. Griffith, "Simultaneous amplification and detection of specific DNA sequences," *Biotechnology*, vol. 10, no. 4, pp. 413–417, April 1992.
[41] C. Ding and C. R. Cantor, "Quantitative analysis of nucleic acids–the last few years of progress," *J Biochem Mol Biol*, vol. 37, no. 1, pp. 1–10, January 2004.
[42] T. Morrison, J. Hurley, J. Garcia, K. Yoder, A. Katz, D. Roberts, J. Cho, T. Kanigan, S. E. Ilyin, D. Horowitz, J. M. Dixon, and C. J. Brenan, "Nanoliter high throughput quantitative PCR," *Nucleic Acids Res*, vol. 34, no. 18, 2006.
[43] G. Kamberova and S. Shah, *DNA Array Image Analysis: Nuts and Bolts (Nuts and Bolts series)*. DNA Press, 2002.
[44] G. Bhanot, Y. Louzoun, J. Zhu, and C. DeLisi, "The importance of thermodynamic equilibrium for high throughput gene expression arrays," *Biophys J*, vol. 84, no. 1, pp. 124–135, January 2003.
[45] J. Bishop, S. Blair, and A. M. Chagovetz, "A competitive kinetic model of nucleic acid surface hybridization in the presence of point mutants," *Biophys. J*, vol. 90, no. 3, pp. 831–840, February 2006.
[46] D. Erickson, D. Li, and U. J. Krull, "Modeling of DNA hybridization kinetics for spatially resolved biochips," *Analytical Biochemistry*, vol. 317, no. 2, pp. 186–200, June 2003.
[47] K. H. Siegmund, U. E. Steiner, and C. Richert, "Chipcheck-a program predicting total hybridization equilibria for DNA binding to small oligonucleotide microarrays," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 2153–2162, November 2003.
[48] T. D. Schneider, "Theory of molecular machines. I. Channel capacity of molecular machines," *J. Theor. Biol.*, vol. 148, no. 1, pp. 83–123, 1991.
[49] M. H. Garzon, K. Bobba, and A. Neel, "Efficiency and reliability of semantic retrieval in DNA-based memories," in *DNA Computing*, 2004, pp. 157–169.
[50] H. Vikalo, B. Hassibi, and A. Hassibi, "A statistical model for microarrays, optimal estimation algorithms, and limits of performance," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 2444–2455, 2006.
[51] K. Pappaert, P. Van Hummelen, J. Vanderhoeven, G. V. Baron, and G. Desmet, "Diffusion-reaction modelling of DNA hybridization kinetics on biochips," *Chemical Engineering Science*, vol. 58, no. 21, pp. 4921–4930, November 2003.
[52] S. A. Tsaftaris, A. K. Katsaggelos, T. N. Pappas, and T. E. Papoutsakis, "DNA-based matching of digital signals," in *Proc. of ICASSP 2004*, vol. 5, 2004, pp. 581–4.

Sotirios A Tsaftaris received the PhD and MSc degrees in Electrical and Computer Engineering from Northwestern University in 2006 and 2003, respectively. He received the diploma in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 2000.

Since 2006 he has been a Research Assistant Professor with the Department of Electrical Engineering and Computer Science at Northwestern University in Evanston, IL. Since 2009 he also holds an appointment with the Department of Radiology, Feinberg School of Medicine. His research interests are bio-molecular computing applications, bioinformatics, medical imaging analysis (MRI in particular), data mining, and digital copyright management.

Aggelos K Katsaggelos received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in EE from the Georgia Tech, in 1981 and 1985, respectively.

In 1985, he joined the Department of EECS at Northwestern University, where he is currently a Professor (past holder of the Ameritech Chair of Information Technology). He is also the Director of the Motorola Center for Seamless Communications, a member of the Academic Affiliate Staff, NorthShore University Health System, and an affiliated faculty at the Department of Linguistics.

He has published extensively (5 books, 160 journal papers, 380 conference papers, 38 book chapters, 14 patents). He is a Fellow of the IEEE (1998), Fellow of SPIE (2009), and the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE ICME Paper Award (2006), and an IEEE ICIP Paper Award (2007). He was a Distinguished Lecturer of the IEEE Signal Processing Society (20072008).