# RETRIEVAL ACCURACY OF VERY LARGE DNA-BASED DATABASES OF DIGITAL SIGNALS

*Sotirios A. Tsaftaris, and Aggelos K. Katsaggelos*

Department of Electrical Engineering and Computer Science, Northwestern University
2145 Sheridan Rd, 60208, Evanston, USA
phone: 1-847-491-7164, fax: 1-847-491-4455, email: {stsaft, aggk}@eecs.northwestern.edu
web: http://ivpl.eecs.northwestern.edu

## ABSTRACT

*In this paper a simulation of single query searches in very large DNA-based databases that are capable of storing and retrieving digital signals is presented. Similarly to the digital domain, a signal-to-noise ratio (SNR) measure to assess the performance of the DNA-based retrieval scheme in terms of database size and source statistics is defined. With approximations, it is shown that the SNR of any finite size DNA-based database is upper bounded by the SNR of an infinitely large one with the same source distribution. Computer simulations are presented to validate the theoretical outcomes.*

## 1. INTRODUCTION

DNA molecules can be used to store digital signals as was first presented in [1]. DNA as a medium is ideal for long term archival of information for rare access with low maintenance cost and high capacity, and furthermore allows for parallel content based retrieval. The signals are encoded using a look-up table that matches signal values, $i$, to fixed length DNA sequences, called words $w_i$. The design of this look-up table, also known as the codeword design problem in DNA computing [2, 3] is of critical importance. The signals are encoded using this look-up table and then DNA sequences are synthesized to form the DNA molecules that form the database. A unique index DNA sequence is attached to the beginning of a DNA encoded signal. An example of such database is shown in Fig. 1. For each database element $S_m$, with concentration $C_m$, the gray part is the index that identifies the data, which are shown as solid black lines. Data are concatenations of DNA words $w_i$, $i = 0, \ldots, N-1$, in the alphabet $A, T, G, C$, each of length $l$.

To retrieve information from the database query molecules are synthesized. Queries can be signal segments of interest. The query signal is encoded using the same look-up table, but the complementary sequence (using the rule $A \leftrightarrow T, G \leftrightarrow C$) is synthesized and introduced in the solution. The query molecules are labeled with a dye that fluoresces at the event of hybridization. The query molecules will hybridize to complementary molecules in the database as seen in Fig. 1 and will fluoresce. This results in a yes/no answer, based on whether the specific signal is present in the database. The amount of fluorescence is directly proportional to the concentration of query-database element molecular complexes.

The question of perfect hybridization (perfect matching) and imperfect hybridization (partial matching in the mean squared error sense) was also addressed in [1]. By modifying the codeword design problem controlled imperfect hybridizations can be performed.

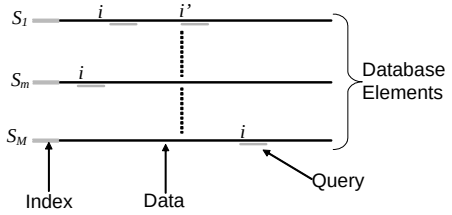


Figure 1: Illustration of hybridizations between query and database elements. (The variable $i$ indicates location.)

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, temperature and salt concentration [4]. Therefore, it is critical to quantify the percentage of fluorescent output that corresponds to desired hybridizations and not to erroneous ones. Consequently, a signal to noise ratio (*SNR*) can be defined, where signal is considered the fluorescence corresponding to desired hybridizations and noise the fluorescent response of undesired ones.

Simulation frameworks to model DNA hybridization interactions were presented in [5, 6]. Hence, concentrations of query-database element complexes can be estimated and *SNR* measurements can be taken. The main contribution of this article is the *SNR* study of querying very large databases with a single query. The presented framework allows for numerical solutions as well as approximations under certain conditions. Following certain approximations, it is shown that the *SNR* of a DNA database is upper bounded by the *SNR* of an infinitely large DNA database with the same source distribution. This result is in agreement with previously presented error measurements [6] which were derived empirically. Consequently, in terms of retrieval accuracy, there is a performance gain as the database size increases.

The framework can also be used to simulate and optimize laboratory protocols such as polymerase chain reaction (PCR), primer and oligo design, microarray oligo design, and microarray simulations.

This paper is organized as follows. In Section 2 the characteristics of a DNA-based database system that can store digital signals are sketched. The framework for kinetic modeling of single query searches in DNA databases and performance evaluation using an *SNR* metric is presented in Section 2. The study on the *SNR* of an infinitely large database is presented in Section 4. Simulation results are given in Section 5. Finally in Section 6 conclusions are given along with possible future extensions and applications in life sciences.

## 2. MODELING SINGLE QUERY RETRIEVAL

The overall system can be described with the following parameters and inputs:

1. $M$ database elements $S_m$ ($M$ digital signals), each of concentration $C_m$ and sequence information $s_m$ each of length $L$, $m = 1,\ldots,M$.
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. Temperature $T$ and salt concentration $|Na^{++}|$.

Queries are concatenations of complements of codewords. It is assumed here that the query is a single codeword $Q = w_i^C$ and that only perfectly aligned linear complexes (only internal mismatches present) are formed, therefore only word to word interactions are considered.

Under these assumptions query database element complexes are actually codeword pairs $w_i^C w_j$. The main objective in this section is to estimate the concentration of complexes $w_i^C w_j$, denoted by $|w_i^C w_j|$, in equilibrium, assuming that all database elements have equal concentration $C_m = C$.

Since there are $M$ database elements and each database element is $L$ bases long, the total number of complexes $N_T$ is equal to $N_T = M\frac{L}{l} = M \cdot k$. For a given $w_i^C$, there are $N$ possible hybridization reactions of the form,

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

that fully describe all possible interactions. The parameters $K_f$ and $K_r$ are called respectively the forward and reverse rate constants. They depend on environmental parameters and laboratory settings and thus are hard to estimate. Therefore, subsequently an equilibrium analysis is sought after.

Under an equilibrium assumption, the differential equations that describe the mass action equations that satisfy Eq. 1 become polynomial equations. Therefore the equilibrium constant $K_{ij}$, can be defined as the constant of the complex $w_i^C w_j$, according to

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

where $|w_i^C|$ and $|w_j|$ are the concentrations of the unhybridized (free) $w_i^C$ and $w_j$, respectively, $\Delta G_{ij}$ 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 can be estimated using the nearest neighbor model in [4].

There exist $N$ integers $j$ $[0, N-1]$ encoded into a DNA codeword $w_j$ with probability $P(w_j)$. Furthermore it is known that there are $N_T = M \cdot k$ occurrences of codewords in $M$ database elements, therefore 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 law dictates that the mass of the reactants (initial concentration of the query) must equal the mass of the products (concentration of unhybridized/free query and hybridized query-word complexes), that is

$$\begin{aligned} |w_i^C|_o &= |w_i^C|_{\text{free}} + |w_i^C|_{\text{hybridized}} \\ &= |w_i^C| + \sum_{j=0}^{N-1} |w_i^C w_j| = |w_i^C| + \sum_{j=0}^{N-1} K_{ij} |w_i^C| \cdot |w_j|. \end{aligned} \tag{4}$$

Similarly, the mass conservation equation for each codeword $w_j$ is given by

$$|w_j|_o = |w_j| + |w_i^C w_j| = |w_j| + K_{ij} |w_i^C| \cdot |w_j|, \tag{5}$$

or

$$|w_j| = \frac{|w_j|_o}{1 + K_{ij} |w_i^C|}. \tag{6}$$

By substituting Eq. 6 into Eq. 4, an equation with the unknown $|w_i^C|$ can be defined:

$$|w_i^C|_o = |w_i^C| + \sum_{j=0}^{N-1} K_{ij} |w_i^C| \cdot \frac{|w_j|_o}{1 + K_{ij} |w_i^C|}. \tag{7}$$

Following the approach in [7] it can be shown that Eq. 7 has a unique solution for $|w_i^C|$ that can be found using the bisection method, which is henceforth denoted by $|w_i^C|_B$. By dividing Eq. 7 by $|w_i^C|_o$ and setting

$$\omega_i = \frac{|w_i^C|}{|w_i^C|_o} \quad \omega_i \in [0,1], \tag{8}$$

the following function can be defined

$$f(\omega_i) = -1 + \omega_i + \sum_{j=0}^{N-1} \frac{|w_j|_o}{|w_i^C|_o} \cdot \frac{\omega_i}{\frac{1}{K_{ij} \cdot |w_i^C|_o} + \omega_i}. \tag{9}$$

$f(\omega_i)$ has a single root since it is a monotonically increasing function as seen bellow

$$f'(\omega_i) = 1 + \sum_{j=0}^{N-1} \frac{|w_j|_o}{|w_i^C|_o} \cdot \frac{\frac{1}{K_{ij} \cdot |w_i^C|_o}}{\left(\frac{1}{K_{ij} \cdot |w_i^C|_o} + \omega_i\right)^2} > 0. \tag{10}$$

In addition since $f(0) = -1 < 0$ and

$$f(1) = \sum_{j=0}^{N-1} \frac{|w_j|_o}{|w_i^C|_o} \cdot \left(\frac{1}{K_{ij} \cdot |w_i^C|_o} + 1\right)^{-1} > 0. \tag{11}$$

then according to the intermediate value theorem there exists a unique $\omega_i^S \in [0,1]$ such that $f(\omega_i^S) = 0$.

To find an approximate solution to $f(\omega_i) = 0$, Eq. 3 is substituted into Eq. 9 to define the following equation

$$\begin{aligned} 1 - \omega_i &= \sum_{j=0}^{N-1} \frac{P(w_j) \cdot M \cdot k \cdot C}{|w_i^C|_o} \cdot \frac{\omega_i}{\frac{1}{K_{ij} \cdot |w_i^C|_o} + \omega_i} \\ &= M \cdot k \cdot \frac{C}{|w_i^C|_o} \cdot \sum_{j=0}^{N-1} P(w_j) \cdot \frac{\omega_i}{\frac{1}{K_{ij} \cdot |w_i^C|_o} + \omega_i}. \end{aligned} \tag{12}$$

If $\rho = \frac{C}{|w_i^C|_o}$ and $h(\omega_i) = \omega_i \left( \frac{1}{K_{ij} \cdot |w_i^C|_o} + \omega_i \right)^{-1}$ are set, the following is obtained

$$\frac{1-\omega_i}{M \cdot k \cdot \rho} = \sum_{j=0}^{N-1} P(w_j) \cdot h(\omega_i). \quad (13)$$

Each $h(\omega_i)$ inside the sum takes values in the range $[0,1)$ for $\omega_i \in [0,1]$. This is due to the fact that the term $\frac{1}{K_{ij} \cdot |w_i^C|_o}$ is much smaller than 1 therefore $h(1) \simeq 1$.

As $\omega_i \to 0$, $h(\omega_i)$ can be approximated by a linear term by finding its first derivative, and evaluating it close to zero to essentially find its tangent at zero as presented in [7]:

$$h(\omega_i) = \frac{\omega_i}{\frac{1}{K_{ij} \cdot |w_i^C|_o} + \omega_i} \approx K_{ij} |w_i^C|_o \cdot \omega_i. \quad (14)$$

Substituting the above equation and $\rho = \frac{C}{|w_i^C|_o}$ in Eq. 13 after some manipulations the following expression is defined

$$\omega_i = \frac{1}{1 + M \cdot k \cdot C \cdot \sum_{j=0}^{N-1} P(w_j) \cdot K_{ij}} = \frac{|w_i^C|}{|w_i^C|_o}, \quad (15)$$

or finally

$$|w_i^C| = \frac{|w_i^C|_o}{1 + M \cdot k \cdot C \cdot \sum_{j=0}^{N-1} \cdot K_{ij} \cdot P(w_j)}. \quad (16)$$

The last equation can be further approximated as

$$|w_i^C| \approx \frac{|w_i^C|_o}{M \cdot k \cdot C \cdot \hat{K}}, \quad (17)$$

where $\sum_{j=0}^{N-1} K_{ij} \cdot P(w_j) = \hat{K}$.

The linear approximation Eq. 15 is close to the real solution $\omega_i^S$, if $(M \cdot k \cdot \rho)^{-1} < min\{P(w_0), P(w_1), ..., P(w_{N-1})\}$ holds. This is equivalent to $|w_i^C|_o < min\{|w_0|_o, |w_1|_o, ..., |w_{N-1}|_o\}$, which is termed thereafter as diluted query concentration.

By substituting Eq. 6 in Eq. 2 the following expression is defined

$$|w_i^C w_j| = K_{ij} \cdot |w_i^C| \cdot |w_j| = \frac{|w_j|_o \cdot K_{ij} \cdot |w_i^C|}{1 + K_{ij} |w_i^C|}. \quad (18)$$

Finally, substituting Eqs. 3 and 17 in the above equation it is obtained

$$|w_i^C w_j| = \frac{M \cdot k \cdot C \cdot P(w_j) \cdot K_{ij} \cdot |w_i^C|_o}{M \cdot k \cdot C \cdot \hat{K} + K_{ij} |w_i^C|_o}. \quad (19)$$

Having estimated $|w_i^C w_j|$ the concentration ratios and the *SNR* is defined next.

## 3. CONCENTRATION RATIOS AND SIGNAL-TO-NOISE RATIO

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

$$\frac{|w_i^C w_j|}{|w_i^C w_{j'}|} = \frac{K_{ij}}{K_{ij'}} \cdot \frac{|w_j|_o}{|w_{j'}|_o} \cdot \frac{1 + K_{ij'} \cdot |w_i^C|}{1 + K_{ij} \cdot |w_i^C|} \quad (20)$$

Utilizing Eqs. 17 and 3 after some simple manipulations the following expression for the ratio can be defined

$$\frac{|w_i^C w_j|}{|w_i^C w_{j'}|} = \frac{K_{ij}}{K_{ij'}} \cdot \frac{P(w_j)}{P(w_{j'})} \cdot \frac{M \cdot k \cdot \frac{C}{|w_i^C|_o} \cdot \hat{K} + K_{ij'}}{M \cdot k \cdot \frac{C}{|w_i^C|_o} \cdot \hat{K} + K_{ij}}. \quad (21)$$

Eq. 21 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 term that highlights the dependency on the ensemble of fragments, through $M \cdot k \cdot \frac{C}{|w_i^C|_o} \cdot \hat{K}$. Tulpan *et al*. [2], hint on this dependency without actually deriving it.

Similarly to [6] the signal-to-noise ratio (*SNR*) of a search with query $w_i^C$ can be defined as:

$$SNR(w_i^C) = \frac{\sum_{j \in \text{ desired}} |w_i^C w_j|}{\sum_{j \in \text{ un-desired}} |w_i^C w_j|}. \quad (22)$$

A new Noise Tolerance Constraint (NTC) was presented in [1] that allows for the design of codewords such that desired hybridizations are those for which $|i-j| \leq T_P$, while un-desired are those for which $|i-j| > T_P$, where $T_P$ is a user supplied parameter that defines the noise tolerance of the codeword set. It has been shown that for single queries codeword sets that follow the NTC, the hybridization strength between codewords is analogous to the mean squared error of the corresponding indices [5].

In this work 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 there can be only codeword pair interactions, desired and un-desired pairs can be explicitly specified. Hence, Eq. 22 can be written as

$$SNR(w_i^C) = \frac{\sum_{j=i-T_P}^{i+T_P} |w_i^C w_j|}{\sum_{j=0}^{i-T_P-1} |w_i^C w_j| + \sum_{j=i+T_P+1}^{N-1} |w_i^C w_j|}, \quad (23)$$

or by dividing with $|w_i^C w_i|$, can be written as

$$SNR(w_i^C) = \frac{1 + \sum_{j=i-T_P}^{i-1} \frac{|w_i^C w_j|}{|w_i^C w_i|} + \sum_{j=i+1}^{i+T_P} \frac{|w_i^C w_j|}{|w_i^C w_i|}}{\sum_{j=0}^{i-T_P-1} \frac{|w_i^C w_j|}{|w_i^C w_i|} + \sum_{j=i+T_P+1}^{N-1} \frac{|w_i^C w_j|}{|w_i^C w_i|}}. \quad (24)$$

The value of $SNR(w_i^C)$ can be calculated by substituting Eq. 20 into the above equation.

## 4. SNR OF AN INFINITELY LARGE DATABASE

In this section expressions for the *SNR* and the error rate of the system as the number of database elements $M$ reaches infinity are derived. Using the assumptions of Section 2, it is shown that a system allowing for a noise tolerant retrieval (like the one used in this work) has a lower error probability than other systems without noise tolerance.

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 Eq. 21 after some basic steps the following can be had

$$\lim_{M\to\infty} \frac{|w_i^C w_j|}{|w_i^C w_i|} = \frac{\infty}{\infty} = \ldots = \frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}. \tag{25}$$

Based on the previous equation, and Eq. 24

$$\begin{aligned}
&\lim_{M\to\infty} SNR(w_i^C) = SNR(w_i^C)_\infty = \\
&= \frac{1 + \sum_{j=i-T_P}^{i-1} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right) + \sum_{j=i+1}^{i+T_P} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right)}{\sum_{j=0}^{i-T_P-1} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right) + \sum_{j=i+T_P+1}^{N-1} \left(\frac{K_{ij}}{K_{ii}} \cdot \frac{P(w_j)}{P(w_i)}\right)} \\
&= \frac{\sum_{j=i-T_P}^{i+T_P} (K_{ij} \cdot P(w_j))}{\sum_{j=0}^{i-T_P-1} (K_{ij} \cdot P(w_j)) + \sum_{j=i+T_P+1}^{N-1} (K_{ij} \cdot P(w_j))},
\end{aligned} \tag{26}$$

or if a uniform distribution for $w_j$ is assumed

$$\lim_{M\to\infty} SNR(w_i^C) = \frac{\sum_{j=i-T_P}^{i+T_P} K_{ij}}{\sum_{j=0}^{i-T_P-1} K_{ij} + \sum_{j=i+T_P+1}^{N-1} K_{ij}}. \tag{27}$$

It will be shown experimentally that Eq. 26 is an upper-bound for the *SNR* performance of a database with a finite number of database elements. That is,

$$SNR(w_i^C)_M \le SNR(w_i^C)_\infty. \tag{28}$$

The corresponding error probability can be defined as

$$E_\infty(w_i^C) = \frac{\sum_{j\in \text{ un-desired}} |w_i^C w_j|}{\sum_{j\in \text{ all}} |w_i^C w_j|} = \frac{1}{1 + SNR(w_i^C)_\infty}. \tag{29}$$

For a uniform distribution of codewords the corresponding error probability is in agreement with the definition of *computational incoherence* in [6] (the probability of error in annealing reactions). However, the analysis in [6] is rather qualitative than quantitative. In this section it was shown (using a linear approximation), that at infinity the error rate is only a function of the source statistics $P(w_j)$ and the equilibrium constants $K_{ij}$.

The performance of the proposed retrieval and codeword design system can be compared with codeword designs that do not allow any error. Assuming a codeword set with equilibrium constants $K'_{ij}$ that does not allow any error during retrieval (i.e., a match is declared only when it is perfect). The *SNR* of a search $w'_i$ at infinity can be found by replacing $K_{ij}$ with $K'_{ij}$, and setting $T_P = 0$ in Eq. 27, that is,

$$\lim_{M\to\infty} SNR(w_i'^C) = \frac{K'_{ii}}{\sum_{j=0}^{i-1} K'_{ij} + \sum_{j=i+1}^{N-1} K'_{ij}}. \tag{30}$$

By comparing Eqs. 30 and 27, in order for the *SNR* expression in Eq. 30 to be larger than or equal to the expression in Eq. 27 either

$$K'_{ii} \ge \sum_{j=i-T_P}^{i+T_P} K_{ij}, \tag{31}$$

or

$$\sum_{j=0}^{i-1} K'_{ij} + \sum_{j=i+1}^{N-1} K'_{ij} \le \sum_{j=0}^{i-T_P-1} K_{ij} + \sum_{j=i+T_P+1}^{N-1} K_{ij}, \tag{32}$$

must hold.

This shows that controlled cross-hybridization is actually beneficial in terms of *SNR* when designing such systems, since Eqs. 31 or 32 need to be satisfied by a system not allowing for any hybridization errors.

## 5. COMPUTER SIMULATION OF DNA DATABASES

In this section simulation results are presented, which were obtained when the models and derivations of the previous section were implemented in a computing language (MATLAB) to simulate data retrieval in a test DNA database. As a signal to DNA encoding strategy the codeword set of 32 words of length 19 derived with the algorithm presented in [3] for $T_P = 3$ is used.

The accuracy of the approximate solution of Eq. 7 provided by Eq. 17 was compared with the computational solution $|w_i^C|_B$. The *SNR* was chosen as a comparison metric. The query was the integer $q_d = \{14\}$. The equilibrium constants between $q_d = \{14\}$ and the signal values $0, \ldots, 31$ were found. A uniform distribution was assumed for the source, that is $P(w_j) = 1/32$, $j = 0, \ldots, 31$. $M$ was equal to 3 and there were $k = 20$ words per database element. The initial concentration of each database element was $C = 10^{-5} {}^{mol}/_{Liter}$. According to Eq. 23 and $T_P = 3$ the *SNR* can be found as

$$SNR(w_{14}^C) = \frac{\sum_{j=11}^{17} |w_{14}^C w_j|}{\sum_{j=0}^{10} |w_{14}^C w_j| + \sum_{j=18}^{31} |w_{14}^C w_j|}. \tag{33}$$

The *SNR* for $\rho = C/|w_{14}^C|_o = 10^{-3}, \ldots, 10^2$ was found using the bisection method and the approximation. The results are shown in Fig. 2. When $\rho > 1$ the approximate solution is very close to the exact computational solution, while for $\rho < 1$ the approximation does not hold. Furthermore, the *SNR* increases as $\rho$ increases, which is a clear indication that

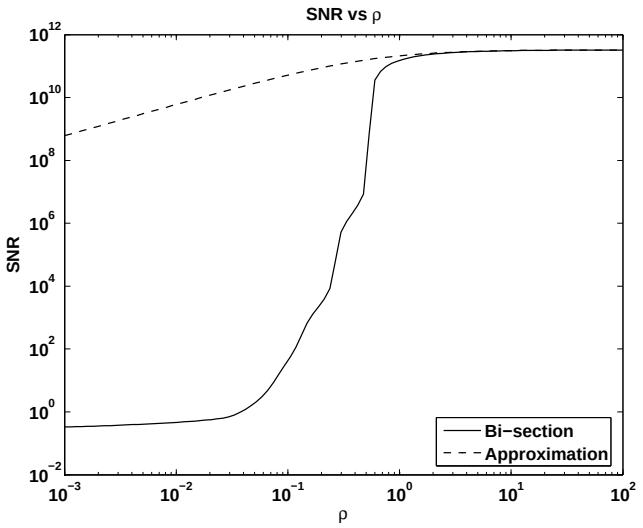


Figure 2: *SNR* for various $\rho$.

competitive hybridization is a critical and desired component for the performance of the proposed system.

To find the *SNR* as the database size increases the above experiment was repeated but the size of the database ($M$) was increased at each iteration. Specifically, the goal was to verify the validity of Eq. 26.

Assuming a uniform distribution Eq. 27 becomes

$$\lim_{M\to\infty} SNR(w_{14}^C) = \frac{\sum_{j=11}^{17} K_{14j}}{\sum_{j=0}^{10} K_{14j} + \sum_{j=18}^{31} K_{14j}}. \qquad (34)$$

In Fig. 3 the *SNR* of a database of size $M$, where $M = 1,\ldots,10^{10}$ for $\rho = 100, 10, 1, 0.1, 0.01$, is shown. With the dashed line the value of $SNR(w_{14}^C)_\infty = 3.2681 \cdot 10^{11}$ is also shown. It can be seen that for large $M$, the bound at infinity can be achieved independently of the value of $\rho$. Furthermore, it is observed that as long as the query concentration is kept in dilute, i.e., $\rho > 1$, the performance of the database is very close to the maximum achievable *SNR*. The graph also hints at an estimate of $\rho$ that should be relative to the database size. As a rule of thumb, therefore, $\rho = 1$ should be adequate to achieve good performance for databases with $M > 100$.

## 6. CONCLUSION

In this paper a framework to simulate single query situations was presented. The kinetic analysis and formulation allows for numerical solutions, as well as approximate solutions under certain conditions. When approximations are utilized, it was shown 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. A number of simulation results were presented that verify and support the claims.

In terms of applications of interest to the life sciences community, the proposed simulation framework can be used to simulate and optimize laboratory protocols such as polymerase chain reaction (PCR), primer and oligo design, microarray oligo design, and microarray simulations. For example, the concentration ratios expression of Eq. 21 can be used to find candidate loci of genes for microarray probe design and PCR primer design.

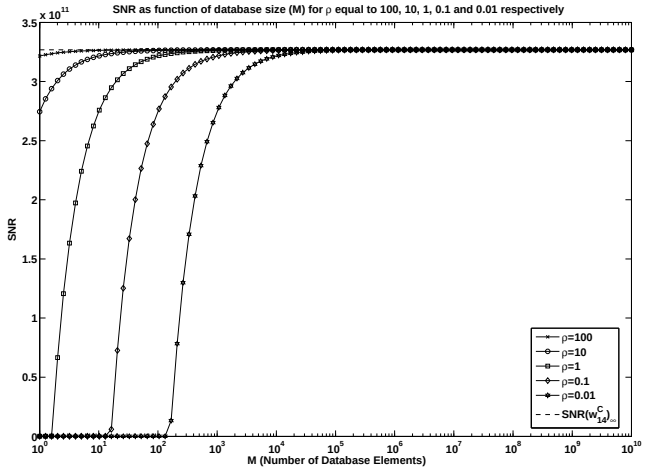


Figure 3: *SNR* as a function of database size $M$, for various $\rho$. The upper bound $SNR(w_{14}^C)_\infty$ is plotted as a dashed line for comparison.

## REFERENCES

[1] 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 Magazine*, vol. 21, no. 6, pp. 57–61, 2004.

[2] 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.

[3] S. A. Tsaftaris and A. K. Katsaggelos, "A new codeword design algorithm for DNA based storage and retrieval of digital signals," in *Preproceedings DNA-based computers 11*, London, Ontario, Canada, 2005.

[4] 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, 1998.

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

[6] 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 *Proc. of the Genetic and Evolutionary Computation Conference*, vol. 2. Orlando, Florida, USA: Morgan Kaufmann, 13-17 July 1999, pp. 1829–1834.

[7] S. A. Tsaftaris, V. Hatzimanikatis, and A. K. Katsaggelos, "*In silico* estimation of annealing specificity of query searches in DNA databases," *J. of the Japan Society of Simulation Technology*, vol. 24, no. 4, pp. 268–276, 2005.